Efficient and Stable Online Learning for Developmental Robots

Von der
Carl-Friedrich-Gauss-Fakultät
der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung des Grades einer

Doktoringenieurin (Dr.-Ing.)

genehmigte Dissertation

von
Rania Rayyes
geboren am 28 Juni 1986
in Latakia

Eingereicht: 16 Oktober 2020
Disputation: 16 Dezember 2020
1. Referent: Prof. Dr. Jochen Steil
2. Referent: Prof. Dr. Jun Tani
3. Referent: Prof. Dr. Jochen Triesch

2020

Thesis advisor: Jochen Steil Rania Rayyes

Efficient and Stable Online Learning for Developmental Robots

Abstract

Recent progress in robotics and cognitive science has inspired a new generation of more versatile robots, so-called developmental robots. Many learning approaches for these robots are inspired by developmental processes and learning mechanisms observed in children. It is widely accepted that developmental robots must autonomously develop, acquire their skills, and cope with unforeseen challenges in unbounded environments through lifelong learning. Continuous online adaptation and intrinsically motivated learning are thus essential capabilities for these robots. However, the high sample-complexity of online learning and intrinsic motivation methods impedes the efficiency and practical feasibility of these methods for lifelong learning. Consequently, the majority of previous work has been demonstrated only in simulation.

This thesis devises new methods and learning schemes to mitigate this problem and to permit direct online training on physical robots. A novel intrinsic motivation method is developed to drive the robot's exploration to efficiently select what to learn. This method combines new knowledge-based and competence-based signals to increase sample-efficiency and to enable lifelong learning.

While developmental robots typically acquire their skills through self-exploration, their autonomous development could be accelerated by additionally learning from humans. Yet there is hardly any research to integrate intrinsic motivation with learning from a teacher. The thesis therefore establishes a new learning scheme to integrate intrinsic motivation with learning from observation.

The underlying exploration mechanism in the proposed learning schemes relies on Goal Babbling as a goal-directed method for learning direct inverse robot models online, from scratch, and in a learning while behaving fashion. Online learning of multiple solutions for redundant robots with this framework was missing. This thesis devises an incremental online associative network to enable simultaneous exploration and solution consolidation and establishes a new technique to stabilize the learning system.

The proposed methods and learning schemes are demonstrated for acquiring reaching skills. Their efficiency, stability, and applicability are benchmarked in simulation and demonstrated on a physical 7-DoF Baxter robot arm.

Keywords: Developmental robots, online learning, intrinsic motivation, learning from observation, sample-efficiency, associative network, lifelong learning.

Zusammenfassung

Jüngste Entwicklungen in der Robotik und den Kognitionswissenschaften haben zu einer Generation von vielseitigen Robotern geführt, die als "Developmental Robots" bezeichnet werden.

Lernverfahren für diese Roboter sind inspiriert von Lernmechanismen, die bei Kindern beobachtet wurden. "Developmental Robots" müssen autonom Fertigkeiten erwerben und unvorhergesehene Herausforderungen in uneingeschränkten Umgebungen durch lebenslanges Lernen meistern. Kontinuierliches Anpassen und Lernen durch intrinsische Motivation sind daher wichtige Eigenschaften.

Allerdings schränkt der hohe Aufwand beim Generieren von Datenpunkten die praktische Nutzbarkeit solcher Verfahren ein. Daher wurde ein Großteil nur in Simulationen demonstriert.

In dieser Arbeit werden daher neue Methoden konzipiert, um dieses Problem zu meistern und ein direktes Online-Training auf realen Robotern zu ermöglichen.

Dazu wird eine neue intrinsisch motivierte Methode entwickelt, die während der Umgebungsexploration effizient auswählt, was gelernt wird.

Sie kombiniert neue wissens- und kompetenzbasierte Signale, um die Sampling-Effizienz zu steigern und lebenslanges Lernen zu ermöglichen.

Während "Developmental Robots" Fertigkeiten durch Selbstexploration erwerben, kann ihre Entwicklung durch Lernen durch Beobachten beschleunigt werden.

Dennoch gibt es kaum Arbeiten, die intrinsische Motivation mit Lernen von

interagierenden Lehrern verbinden.

Die vorliegende Arbeit entwickelt ein neues Lernschema, das diese Verbindung schafft.

Der in den vorgeschlagenen Lernmethoden genutzte Explorationsmechanismus beruht auf Goal Babbling, einer zielgerichteten Methode zum Lernen inverser Modelle, die online-fähig ist, kein Vorwissen benötigt und Lernen während der Ausführung von Bewegungen ermöglicht.

Das Online-Lernen mehrerer Lösungen inverser Modelle redundanter Roboter mit Goal Babbling wurde bisher nicht erforscht. In dieser Arbeit wird dazu ein inkrementell lernendes, assoziatives neuronales Netz entwickelt und eine Methode konzipiert, die es stabilisiert.

Das Netz ermöglicht deren gleichzeitige Exploration und Konsolidierung.

Die vorgeschlagenen Verfahren werden für das Greifen nach Objekten demonstriert. Ihre Effizienz, Stabilität und Anwendbarkeit werden simulativ verglichen und mit einem Roboter mit sieben Gelenken demonstriert.

Stichworte: Developmental robots, Online-Lernen, intrinsische Motivation, Lernen durch Beobachten, Sampling-Effizienz, assoziative Netze, lebenslanges Lernen.

Contents

List of figures

List of Tables

Acknowledgments

First and most, I would like to deeply thank my great supervisor Prof. Jochen Steil for all his fruitful help, continuous support, valuable advice, and guidance to complete my Ph.D. successfully. I am very grateful for his endurable impact on my research and life. Prof. Steil helped me to grow and build up my future career with all his support and confidence. The friendly atmosphere he assures in the group makes us all feel home.

I would like to extend my gratitude to Dr. Michael Spranger for the great internship opportunity to spend five months at Sony Computer Science Lab in Tokyo, and for his fruitful cooperation and advice. This project was a big addition to my dissertation.

Furthermore, I would like to thank all my coauthors, my previous and current colleagues at CoR-Lab in Bielefeld, IRP in Braunschweig, and Sony CSL for all the fruitful discussions, help, and support.

I cannot forget to thank DAAD (Deutscher Akademischer Austauschdienst), specially my DAAD advisor Frau. Birgit Klaes.

I would like also to thank my reviewers Prof. Jochen Triesch and Prof. Jun Tani for their time, consideration, and evaluations.

I would not be where I am now without the endless support of my family.

This project has been funded by DAAD scholarship ”Research Grants – Doctoral Programme in Germany” and the Institute for Robotics and Process Control (IRP). My internship project has been funded by Sony Computer Science Lab.

How can we devise efficient and stable online data-driven learning methods for developmental robots with direct online training on real robots?

1 Introduction

1.1 Motivation: The Challenges in Lifelong Learning for Developmental Robots

Developmental robotics is a highly interdisciplinary research field which devises new approaches to robotics inspired by developmental principles and learning mechanisms observed in children [1–7]. It aims to build more versatile and adaptive robots by linking natural and artificial systems [1, 2, 8]. It has also gained a lot of attention in cognitive science and developmental psychology [1–3], as it provides computational models and experimental platforms for a better understanding of biological development.

Developmental robots must autonomously develop and adapt in open environments throughout their life-time which is referred to as lifelong learning [5, 8, 9].

One of the fundamental tasks for these robots is learning sensorimotor skills and motor coordination, e.g., reaching [4, 8, 10]. In contrast to the classical industrial robots, which accomplish repetitive predefined tasks, developmental robots must solve unpredictable tasks and cope with unforeseen challenges. Developmental robots must learn new skills unspecified at design time through self-exploration [11]. They must also adapt to time-dependent changes in the environment and robot dynamics (e.g., friction [12], tool usage [13, 14], changing scenery [15, 16]), and autonomously explore their environment. Computational models for intrinsic motivation [17–20] have shown a great potential to tackle these challenges through driving the robot by internally generated signals to select what to learn and where to explore in an open-ended (i.e., unbounded) environment.

Although developmental robots should acquire their skills through self-exploration and intrinsic motivation, they often share the environment with humans and it is beneficial to also learn from these humans. Inferring their goals and imitating their behavior can shift the focus of the robot toward new areas of the workspace to discover, as well as toward important tasks and novel outcomes to learn. Hence, the robot's exploration, which is autonomously guided by intrinsic motivation, could be additionally guided by observing human demonstrations in order to accelerate the autonomous development of these robots. Therefore, intrinsic motivation methods have a great potential to be integrated with "learning from a teacher" methods (i.e., learning from demonstration [11], imitation learning [19], and learning from observation [21])), with the aforementioned remarkable advantages for developmental robots. Yet, there is hardly any active research to integrate these two fields. It is worth mentioning that learning from observation seems the most compatible with human learning [22, 23] since it does not require strictly copying the teacher's behavior but rather achieving similar outcomes. Humans and robots have fundamentally different models, e.g., different kinematics and dynamics models. This favors observational learning for developmental robots over the other "learning from a teacher" methods.

Devising *efficient online* learning methods and schemes for developmental robots is very challenging, specifically on real robots, since data acquisition is typically costly in terms of time, wear, and tear.

First, these methods need to handle lifelong learning in unbounded environments and have to assure scalability to high degrees of freedom (DoF). Therefore, a major challenge in this field is the high sample-complexity of the proposed methods [8, 24, 25], i.e., the dense sampling required to approximate the learned function with reasonable accuracy. This can be only partially mediated by intrinsic motivation methods since these methods are themselves mostly data-driven [24, 25]. Therefore, most previous methods have been demonstrated only in simulation as a proof of concept, and only a few on real robots (e.g., [25–29]).

Second, *online* learning is an essential capability providing *adaptation* in lifelong learning. The online learning in this thesis refers to the ability of online sampling and performing a learning step after each generated sample. The learners should be updated on the fly to be tailored to the learned problem. In contrast, offline learning overly restricts robot adaptation. It needs to retrain the learner in case of changes in the environment or robot dynamics. Offline learning also requires full data sets to be stored to train the model offline, which is not feasible in lifelong learning that has to strictly limit the storage of experience. Yet, all online model learning methods tend to require a lot of training samples - especially in model learning and in robot kinematics and dynamics learning tasks - and the number of training samples increases exponentially with the number of "DoF" [12, 30]. This imposes additional difficulties for *direct online training* on real robots. It is therefore a persistent research question on how to devise efficient online intrinsic motivation methods for real robot applications.

Third, online data-driven learning requires incremental online learners to approximate the models online, i.e., the complexity of the learning model should be *adapted* continuously and cannot be set in advance to the yet unknown learning problem. However, the construction of dynamic neural networks, for example, incrementally from scratch and with continuous online update has potential *stability* problems in the presence of noisy data in real robot applications [31].

Fourth, combining intrinsic motivation methods with "learning from a teacher" methods imposes additional challenges for developmental robots. The robot then must have an *exploration strategy* to decide how to explore, i.e., whether the robot is driven by intrinsic motivation or learning from humans. The robot must also manage the *exploitation-exploration trade-off* and decide autonomously when to explore, i.e., whether to explore further at the cost of additional training or to exploit its prior-knowledge/current knowledge in order to achieve similar outcomes as humans. Therefore, another research question is how to integrate intrinsic motivation with learning from observation and enable the robot to decide autonomously how and when to explore.

Finally, considering high DoF redundant robots, e.g., humanoids [1], brings up the questions of which solution (configuration) to select and learn to achieve a required task, and whether it is possible to learn multiple solutions *online* for each task. The underlying exploration method in this thesis relies on Goal Babbling [32] to leverage its desirable advantages for lifelong learning. Goal Babbling permits direct inverse model learning, online, from scratch, and in a learning while behaving fashion. This method is inspired by how infants learn their motor skills [33]. The adaptability to various applications and the scalability to many DoFs of this method have been demonstrated in different domains (e.g.,[11, 34–38]) and in real robot applications (e.g., [39]). Goal Babbling has been originally proposed for direct learning of inverse kinematics [32]. On the one hand, Goal Babbling solves redundancy resolution by learning one preferable solution based on the initial robot state [32]. On the other hand, this restricts the *flexibility* of the robots. The versatility of the human learning system, one the contrary, allows learning several solutions to accomplish the required tasks flexibly, e.g., reaching an object with different configurations. [40] combined Goal Babbling with associative dynamic networks [40, 41] to learn multiple solutions of inverse kinematics. However, in [40], the exploration with Goal Babbling has been done first for each solution, and then the solutions are consolidated offline in the network. Still, offline learning is undesirable for lifelong learning as discussed before. Furthermore, stabilizing these dynamic networks is very challenging [31, 42]. So how can Goal Babbling learn *multiple solutions* of inverse models for redundant robots *online* and in a *stable* fashion?

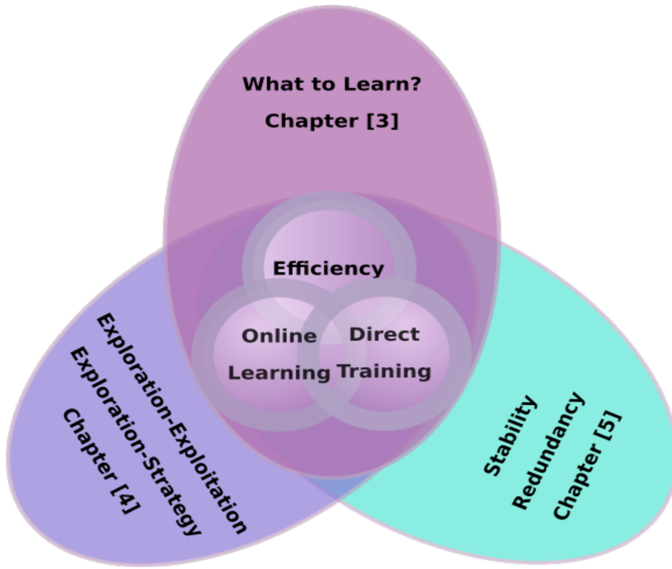


Figure 1.1.1: Lifelong learning challenges for developmental robots: sample-efficiency, direct online training on real robots, fully *online* learning without any offline or batch learning, selecting what to learn, exploration-exploitation trade-off, designing a proper exploration strategy, the stability of the learning system, and learning inverse models with multiple solutions for redundant robots

Despite the interesting proposed previous approaches and the successive achievements, the discussed challenges still prevent full applications of direct online training in real-world scenarios. Moreover, there is hardly any research to integrate intrinsic motivation with learning from observation. These two shortcomings have motivated this dissertation.

1.2 Main Contributions and Goal of the Thesis

The main goal of this thesis is to devise an *efficient and stable online learning scheme for developmental robots with direct online training on real robots*. From the challenges discussed in the previous section for devising such learning schemes (cf. Fig. 1.1.1), three major research questions arise:

1. How can we devise an efficient online intrinsic motivation method which

permits direct online training on physical robots?

2. How can we integrate intrinsic motivation with learning from observation and enable the robot to autonomously decide when and how to explore?

3. How can Goal Babbling learn inverse models, e.g., inverse kinematics, with multiple solutions for redundant robots online and in a stable fashion?

To address these questions, this thesis makes the following main contributions (cf. Fig. 1.2.1):

1. A novel intrinsic motivation method named ”interest measurement” is established in order to increase *sample-efficiency* and intrinsically drive the robot to select *what to learn*. A new knowledge-based and a new competence-based intrinsic motivation signal are devised and combined in the interest measurement method. The knowledge-based signal named ”relative error” enables the robot to select the most informative tasks to learn from and generalize on simpler ones, in order to minimize the number of required samples to learn the models with reasonable accuracy, i.e., increasing *sample-efficiency*. The competence-based signal named ”forgetting factor” enables lifelong learning which requires accommodating new knowledge while retaining the previously gained progress [43] by driving the robot's interest also toward forgotten previously learned tasks. Moreover, an online mental replay method is devised to intensify the robot's experiences in real-world applications.

2. A novel extrinsic-intrinsic learning scheme that combines intrinsic motivation with learning from observation is designed. It allows the robot to explore driven by its interest (intrinsic motivation) as well as guided by observing human demonstrations' outcomes (extrinsic motivation). Extrinsic motivation in this thesis means that the robot is motivated to learn by observing human demonstrations in order to achieve similar outcomes. Three new methods are devised to establish observational learning: Novelty detection, novelty degree, a probabilistic goal selection strategy. The novelty

methods enable the robot to autonomously decide whether to *explore* further or to *exploit* its gained knowledge in order to achieve similar outcomes as humans. They also enable the robot to autonomously decide which *exploration strategy* it must follow: intrinsic or extrinsic motivation learning. The probabilistic goal selection strategy increases the *sample-efficiency* by selecting the most novel observed outcomes to learn from.

3. A novel associative dynamic network called "Online Associative Radial Basis Function network (OARBF)" is devised to enable Goal Babbling to learn multiple solutions of inverse models *online* for redundant robots with simultaneous exploration and solution consolidation. It is constructed incrementally from scratch and updated on the fly to be tailored to the learned problem. A parameter-sharing technique is established to stabilize the network learning dynamics by synchronizing two learners online (incremental regression Local Linear Map (LLM) [44] and OARBF) and leveraging their advantages: stability, accuracy, and multi-model representations. This technique increases the *sample-efficiency* by drastically reducing the number of required samples for OARBF to learn a model with reasonable accuracy. It also speeds up learning by drastically reducing the dimensionality of the parameter space and by synchronizing the learners' update.

The proposed methods are rather general and could be implemented for different learning scenarios. In this thesis, they are demonstrated for motor learning when good body coordination is not yet established. More specifically, the task is to establish motor coordination to acquire basic reaching skills in a learning while behaving fashion, online, and from scratch without any prior-knowledge of the kinematics models.

Four novel learning schemes are established employing some of the proposed methods to provide the robot with different functionalities based on the required task. All the proposed methods and learning schemes in this thesis are fully *online*, updated continuously and instantaneously, and demonstrated with *direct online*

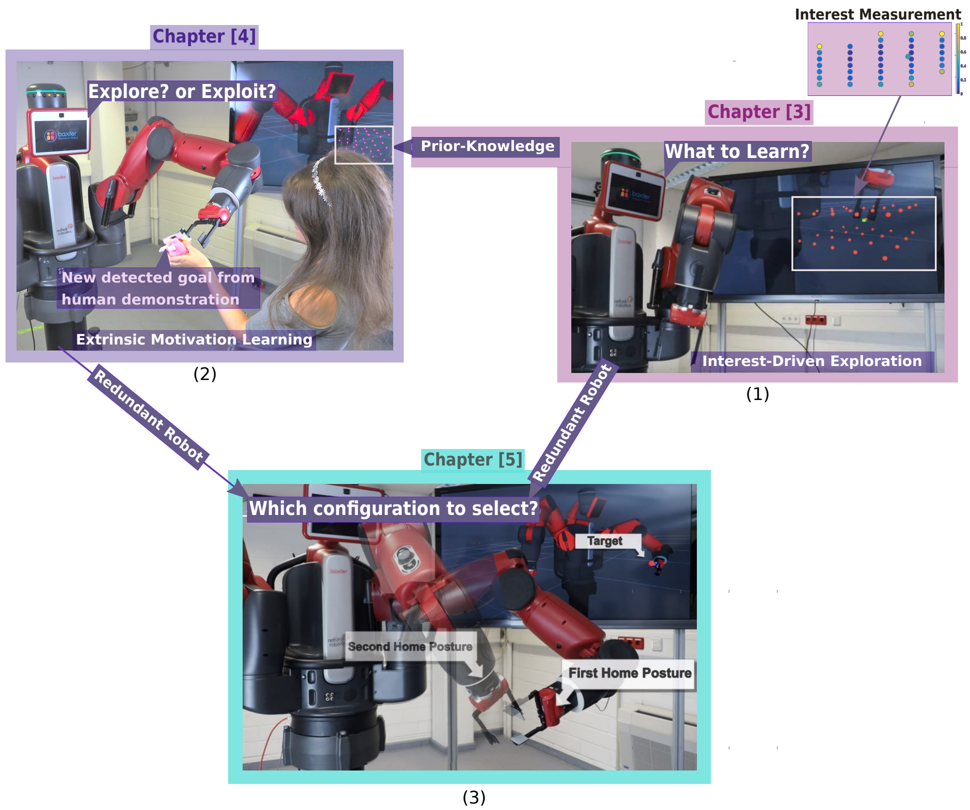


Figure 1.2.1: There are three main contributions in this thesis for devising efficient and stable online learning schemes for developmental robots: (1) a novel intrinsic motivation method named "interest measurement" is devised in chapter 3 to drive the robot to efficiently select what to learn. (2) an extrinsic-intrinsic motivation learning scheme is designed in chapter 4 to integrate intrinsic motivation with learning from observation. It enables the robot to learn from observation and to autonomously decide whether it explores further to achieve the required task, or it exploits its gained prior-knowledge during the intrinsic motivation learning. (3) an incremental online associative dynamic network is devised in chapter 5 to learn different solutions (configurations) with Goal Babbling online for each required task. The robot autonomously selects one of the learned solutions based on its previous state to provide smooth motion

training on a physical 7-DoF robot arm (the left arm of a real Baxter robot by Rethink Robotics [45]). Baxter has a positioning accuracy of 5 *mm* only [45] and the learning while behaving additionally produces very noisy data, which makes the task very challenging on this robot and constitutes an additional stability challenge for the incremental associative network.

The results highlight clearly the scientific progress gained by the proposed methods: (i) increasing efficiency in terms of reducing the number of samples required to learn the model with reasonable accuracy and selecting the most informative samples to learn from; (ii) ensuring fully *online* data-driven learning in terms of instantaneous processing of each received sample and updating the learners as well as all measures on the fly without any intermediate offline learning; (iii) emphasizing adaptability with continuous online update, i.e., the learners are tailored to the learning problem and the robot can quickly adapt to expand its knowledge and learn similar outcomes as humans; (iv) emphasizing stability of the learning system; (v) facilitating direct online training on real robots; (vi) providing suitable exploration strategy to autonomously decide when and how to explore; (vii) emphasizing flexibility, i.e., learning multiple solutions with Goal Babbling for each task online.

1.3 Outline

The thesis is organized as follows: Chapter 2 reviews related work on intrinsic motivation, mental replay, "learning from a teacher" methods, extrinsic motivation, Goal Babbling, and associative dynamic networks.

Chapter 3 devises the interest measurement and integrates it into a novel interest-driven exploration scheme to permit learning while behaving. The chapter then compares the proposed interest measurement with the state-of-the-art to highlight its advantages. A novel online mental replay method is proposed to intensify the robot's experiences. Finally, the full learning scheme is demonstrated in a real robot experiment.

Chapter 4 designs the general structure of the extrinsic-intrinsic motivation learning scheme which combines learning from observation with intrinsic motivation. The chapter then explains the new extrinsic observational learning method in detail. Finally, it evaluates the framework first in an illustrative robot example and then demonstrates it in a real robot experiment.

In chapter 5, OARBF and the parameter-sharing technique are established. The chapter then compares OARBF with the offline version to highlight its advantages. The proposed methods are integrated into the interest-driven exploration scheme and demonstrated in a real robot experiment.

Finally, chapter 6 reviews the main contributions of this thesis as well as the obtained results. It also gives a future outlook on potential extensions of the work.

Learn from the past to build up the future

2
Background

This thesis is primarily related to intrinsic motivation learning, mental replay, integrating intrinsic motivation with "learning from a teacher", and goal-directed exploration and learning (mainly Goal Babbling). It also has some connections to associative dynamic networks and the concept of extrinsic motivation.

One of the fundamental tasks for developmental robots, which is considered in this thesis, is to learn sensorimotor skills and motor coordination from scratch through exploration. Therefore, I only consider data-driven learning methods in my review.

2.1 Intrinsic Motivation

It has been emphasized for a long time that practical, real-world applications of developmental robots require lifelong and online learning. A major challenge in this field is the high sample-complexity of algorithms, which has led to the development of intrinsic motivation approaches to render learning more efficient. Intrinsic motivation in robotics, e.g., [18–20] has been inspired by developmental psychology, where curiosity-driven behavior has been observed in children. For example, children try to discover new things and get easily bored by already known items [20]. Moreover, infants, as well as adults, are engaged in novel activities out of curiosity [46] to improve their knowledge or skills, which has also been shown experimentally in [47].

Intrinsic motivation has been an active research topic for developmental robots [5]. It is used to guide the self-exploration of these robots by internally generated signals in order to actively select what to learn in open-ended environments [18]. Intrinsic motivation approaches are divided into two main categories in the literature [10, 48, 49]: (i) knowledge-based, where the intrinsic motivation signal is generated based on the error between the robot's prediction and the real outcome; (ii) competence-based, where the intrinsic motivation signal is generated based on the learning progress of the robot. However, [47] showed experimentally that humans tend to learn by maximizing their knowledge about the task as well as their competence. Still, there is hardly any intrinsic motivation method to combine knowledge-based with competence-based signals. This thesis proposes a new intrinsic motivation method to combine a new knowledge-based and a new competence-based signal. It shows that this combination yields more efficient goal selection and facilitation of lifelong learning.

2.1.1 Knowledge-Based Intrinsic Motivation

Knowledge-based intrinsic motivation has been considered in different learning contexts with different definitions [10]. In [50], it is differentiated between two types of knowledge-based intrinsic motivation methods: novelty-based and prediction-based, which was highlighted recently also in [17].

On the one hand, some authors considered knowledge-based intrinsic motivation as learning from novel information. The intrinsic motivation signal then was mainly derived based on comparing the newly acquired knowledge with the previous one [10, 17]. For example, the novelty signal in [15] was devised based on the difference between two observed scenes in order to guide the robot to search for more novel scenes to learn from. The learning signal in [51] tried to maximize the diversity of behaviors. In [24], the information gain was maximized by comparing (action-state) distribution before and after learning update. Novelty could be also detected based on a specific error threshold [48]. However, tuning such a threshold for every task in practice is problematic.

On the other hand, knowledge-based intrinsic motivation was also considered as learning from prediction errors [10]. High prediction errors indicate a good learning opportunity. [52] considered the generated reconstruction error signal as an instantaneous reward, i.e., intrinsic motivation signal. The intrinsic reward was proposed in [53] as proportional to the prediction error for each salient event. Furthermore, [54] combined learning from novel situations with learning from high prediction errors. Surprise was defined as a prediction-based method [48, 50]. [50] differentiated between surprise and novelty, and defined surprise as highly unpredictable events to happen. [27] used a dynamics-based surprise signal as a penalty signal to avoid applying high forces while the robot touching an object. Bayesian surprise was used as a curiosity reward in [55].

Minimizing surprise has also been formulated as a free energy principle [56] to explain exploration-exploitation in lifelong learning. The free energy principle assumes that humans try to minimize the long-term average of surprise. Minimizing surprise leads to maximizing model-evidence for intrinsically motivated agents in the context of decision-making. However, the free energy principle cannot explain all intrinsic motivation models, e.g., it cannot describes the approach in [27] where the agent seeks for surprising situations to learn from. Similarly, active inference also refers also to minimizing surprise which is considered as a negative model evidence [57]. Free energy minimization has been implemented, for example, using variational recurrent networks RNN in [58].

[59] investigated experimentally the difference between prediction-based and novelty-based signals to drive the eye movements of humans. The results showed

that novelty-based signals were more effective to drive the learning and suggested that novelty-based intrinsic motivation mechanisms might operate even at an unconscious level.

Still, there is no clear border between these two categories [46], since high errors also indicate novel situations to learn from, as shown later in the novelty detection method in this thesis. High prediction error of events was also considered as a novelty-based signal in [48, 60].

2.1.2 Competence-Based Intrinsic Motivation

Monitoring the average progress of the prediction error was introduced as a competence-based method [61]. [18] considered the performance error over a sliding window of the last n measures and assigned the highest interest to the regions with the highest error changes, regardless whether the error increases or decreases. [19] considered only when the error decreases over the sliding window, i.e., when the robot learns. This approach can avoid unlearnable and unreachable goals for example [19].

The recent works have mostly implemented competence-based methods, e.g., [11, 13, 18–20, 49], where [49] showed that competence-based signals often lead to better performance than knowledge-based ones in learning several reaching tasks with a simulated robot arm. If and how these results transfer to more complex and real-world scenarios is currently an open question.

Novelty in this thesis: While novelty-based approaches have been proposed and implemented as intrinsic motivation methods to drive the learning, the novelty method in this thesis is proposed for extrinsic motivation learning. In addition, the previous novelty-based methods mostly detect novelty by comparing the newly acquired knowledge with the previous one [10, 17], while the novelty method in this thesis detects and measures novelty based only on the current robot's knowledge. In contrast to [48], the error threshold to detect novelty is automatically inferred from the current robot's knowledge.

I follow statics-based methods to determine novelty. They mainly consider novelty as an outlier, meaning that it is significantly different from other samples. For instance, the observed data has different distribution than the already observed ones (e.g., Gaussian distribution), the observed data belongs to unknown classes, or samples fall in regions of low estimated density [46, 50, 62, 63]. However, these approaches potentially misclassify noise as novel data. In this thesis, the novelty is not determined from the distribution of collected samples, but rather from the robot's actual performance error which reflects the robot's actual knowledge. High performance errors are detected as outliers in the performance error distribution in order to indicate novel goals/tasks.

Intrinsic Motivation in Real Applications

The majority of the previous work on intrinsic motivation has been demonstrated only in simulation due to the high sample-complexity of the intrinsic motivation methods. Few works have been demonstrated in real robot applications, e.g., [25–29, 64]. Still, not all of these works have demonstrated efficient learning. For example, [28] demonstrated very slow and inefficient learning. In addition, the robot preformed random movements during the exploration.

Notably high sample-efficiency has been demonstrated in [25, 27]. These methods integrate intrinsic motivation with mental replay (cf. Sec. 2.2) in order to drastically reduce the number of training samples.

One of the limitations of [25] is that it implements a recurrent neural network with a fixed size to learn a limited set of trajectories. The network therefore cannot be adapted to increasing task complexity or scale to a larger workspace in lifelong learning. In contrast, the learners in this thesis are constructed incrementally and their size is updated on the fly to adapt to the current learning problem at hand.

The intrinsic motivation signals in [27] lack continuous online update. The signals are updated once every several training iterations in order to assure learning stability. Besides, the size of the networks is prefixed in [27] similar to [25]. The networks in [27] were trained offline as well as with batch learning. In contrast, the proposed intrinsic motivation in this thesis is updated on the fly at each learning step and demonstrated high stability even in real robot applications.

2.2 Mental Replay

Mental replay is regarded to be an essential component in human learning [65]. Consequently, several replay mechanisms have been proposed for artificial agents as efficient tools to reduce sampling complexity and to speed up the learning process, e.g., Experience Replay [66, 67], Imaginary Experience Replay [68], Hindsight Experience Replay [69, 70], and Mental Replay [25]. The idea of these methods is to store collected training samples in a replay buffer and sample from them again to update the learners frequently. This facilitates deploying data-driven learning methods on real robots since sampling in real robot applications is very costly regarding time and hardware.

Experience Replay [66, 67] stores the full data set and randomly samples again from it to perform mini-batch learning. A similar approach with uniform sampling instead of random sampling was proposed in [71]. This is, however, incompatible with online lifelong learning. [27] sampled mini-batches from a replay buffer containing 1 million transitions. Imaginary Experience Replay [68] and Hindsight Experience Replay [69, 70] use sample augmentation in order to sample imaginary goals. However, this needs the full goal space to be known in advance, which is difficult to be estimated in an open-ended environment. Some approaches extended Hindsight Experience Replay, e.g., [72] which adaptively selects failed experiences for replay based on a curiosity measure, and Prioritized Hindsight Experiences [73] which utilizes an energy function to select which trajectory to be replayed. These are very similar to Prioritized Experience Replay [74]. The latter method prioritizes the most ”surprising” samples with high expected learning progress and replays them more frequently. Mental Replay in [25] was proposed to deal only with spiking neural networks. It generates additional training samples from each trajectory by exploiting the stochastic nature of the spiking network for encoding such trajectories.

Dynamic experience replay [75] stores all data from human demonstration and successful episodic training of the agent. There are several different buffers to randomly sample from, including augmented data from human demonstrations. However, this method is suitable and implemented for offline learning. Invariant Transform Experience Replay [76] uses sample augmentation to generate symmetric trajectories as well as goal augmentation as in [69]. However, the goal augmentation results in samples with a specific error threshold which affects the learning accuracy. Besides, it requires additional geometric information to produce symmetries.

This thesis proposes a new online mental replay method which needs neither data augmentation nor storing full data sets, and it is thus more applicable for lifelong learning applications.

It is worth mentioning that mental replay has been also proposed from a different point of view, as a tool to overcome catastrophic forgetting during incremental learning, e.g., [77, 78].

2.3 Intrinsic Motivation with Learning from a Teacher

While developmental robots must acquire their skills through intrinsically-driven self-exploration, they are also supposed to share the environment with humans [8, 11]. It is, therefore, beneficial for the robot to infer humans' goals and to imitate their behaviors in order to accelerate its autonomous development. Similarly, children and humans explore and learn on their own with the ability to benefit from other teachers or social cues [79].

The previous related works on how artificial agents can learn to perform a task from teacher/expert demonstrations can roughly be categorized in: (i) learning from demonstration, where policy, action, and state information of the teacher/expert is available (e.g.,[11]); (ii) imitation learning, where the (action-state) pairs are available from the teacher/expert demonstration without having access to the policy information (e.g., [80]); (iii) learning from observation, where only the states or the outcomes of the teacher/expert are available (e.g.,[21, 81]).

From a developmental point of view, the two main ways of replicating a movement in infants are imitation and emulation, i.e., learning from observation [82, 83]. [22, 23] pointed out the importance of social learning and observational learning, and highlighted the difference between imitation learning and observational learning, as the latter does not require strictly to copy the teacher's behavior but rather to learn similar behavior, which means exploiting the internal model of the imitator to achieve similar outcomes. Learning from observation also seems the most compatible with human learning, as humans have neither access to the action information of the demonstrators of their peers nor access to their exact internal models. Similarly, humans and robots do not have the same exact internal models, which makes observational learning the most compatible way for robots to learn from humans. Still, there is hardly any research to integrate intrinsic motivation with learning from observation.

[84] studied the connection between intrinsic motivation and imitation learning for efficient coding in active perception. According to [84], the early stage of intrinsic motivation learning facilitates later the imitation learning when new actions and outcomes are considered. When the agent imitates a new action, the intrinsic signal is generated as a feedback signal to reveal how well the new data is encoded in the sensory system [84]. This study has been discussed in the context of human language learning as well as bird song learning.

There are only a very few works in the literature which combine intrinsic motivation with "learning from a teacher" methods. However, not all of them consider lifelong learning for developmental robots. For example, the work in [85] considered only very limited discrete actions/outcomes. The robot arbitrarily explores possible actions (open/close a box and switch on/off a button) and then is driven by its intrinsic motivation based on the novelty of these discrete limited goal set. The human guidance is expressed by pointing out actions or objects from this set. Hence, the human interaction helps only to shape the exploration in the already specified actions/outcomes and does not lead to discovering novel outcomes/goals. It is unclear how this learning system can generalize to other learning scenarios, e.g., continuous space and open-ended environments.

The main notable work for developmental robots and lifelong learning, up to my knowledge, is [11]. [11] combined intrinsic motivation with learning from demonstration where policy and action information of the teacher is required. However, this information is not always available, and even if so, it does not directly transfer as humans and robots have different internal models. In addition, [11] combined random goal selection with intrinsic motivation. The empirical evaluation of the method was demonstrated in [29]. The approach was also extended in [86] for learning sequences of actions.

This thesis proposes a novel learning scheme which combines intrinsic motivation with learning from observation. To this aim, it develops a new extrinsic motivation observational learning method. It provides an exploration strategy which allows the robot to autonomously decide when and how to explore and efficiently select what to learn. It also allows the robot to autonomously manage the exploration-exploitation trade-off. The novelty threshold is automatically detected and inferred from the robot's knowledge.

2.4 Extrinsic Motivation

Extrinsic motivation in robotics in the literature has several controversial definitions [48], where mostly an extrinsic motivation is considered when the robot receives an external reward [87], e.g., from the environment [48, 88], or external signals, i.e., signals from outside the robot controller [89]. In contrast, the reward in the extrinsic motivation in this thesis is an internal reward to maximize the robot's knowledge motivated by external observations.

The extrinsic motivation in this thesis is inspired by a developmental psychology study [46, 88, 90]: once a human observes a behavioral event, a stimulus will trigger a brain activity, the brain interprets this event as a *novel* event, and slowly gets used to it until this event is no longer novel, this is also known as habituation [46]. Accordingly, when the robot observes a teacher demonstration to learn a new task, the robot is motivated to learn to achieve a similar outcome. Achieving users' goals has been also considered as an extrinsic motivation in [17].

2.5 Goal-Directed Learning and Goal Babbling

It is widely accepted since the 1990's that the human motor control is organized on the basis of forward and inverse models (i.e., the relation between actions/motor commands and outcomes) [91]. Forward models convert actions to outcomes, while inverse models estimate the required motor command to achieve a desired outcome. Learning motor capacities and skills has always been a core topic of developmental robots [1], as mastering the body is fundamental for any embodied agent. Learning forward models under the notion of motor babbling [92] has been proposed first to learn the robot functionality by a random exploration of motor commands [93]. This appears unrealistic, however, for robots with many degrees of freedom. The respective high-dimensional spaces for motor commands cannot be exhausted through exploration randomly or systematically because of a combinatorial explosion. In addition, there is evidence from infant studies that neonates demonstrate goal-directed motion already a few days after birth [33, 94]. For example, they learn how to reach by trying to reach, and they adapt their motion by iterating their tries [33]. These insights motivated researchers to turn to the idea of goal-directed inverse model learning, e.g., [95, 96], where the notation Goal Babbling [95] has been coined. Such models have to deal with the problem of redundancy, which is the problem that a redundant robot has many possible ways to achieve a goal and needs to make a selection from them. And they need to assure the scalability to high dimensions.

Goal Babbling has been proposed in [95] as an efficient means for direct inverse model learning and online bootstrapping of sensorimotor skills [32]. It has originally been proposed for learning inverse kinematics (IK) [95], i.e., learning the required configurations to achieve desired spatial goal positions. It permits incremental online learning, from scratch, and in a learning while behaving fashion [32] inspired by infants' learning of their motor skills [33]. It has already demonstrated strong scalability for high DoF robots, e.g., for 50-DoFs of a planar manipulator [32], 9-DoFs of a soft robot [39], and 9-DoFs of a humanoid [36]. The main advantage of Goal Babbling is that it reduces the search space by exploring the low-dimensional space of goals, e.g. spatial positions in the task space. While a redundant manipulator can achieve the same end-effector position with many possible configurations, Goal Babbling focuses on exploring goal positions rather than exploring motor actions. This increases the diversity of the reached positions, assures the efficiency, and speeds up the learning, as demonstrated with a physical 9-DoF elephant trunk robot [39].

Goal Babbling has been replicated and extended in different learning domains (e.g., socially guided [11], acoustics [34], body orientation [97], learning static forces [35], skill Babbling [98], multi-stage Goal Babbling for simultaneous model learning [37], goal-directed learning of hand-eye coordination [99], tool usage [13, 14], and body model learning through self-touch [38]). Learning inverse models of robot manipulators has been also considered as a promising alternative solution to analytical methods since obtaining an accurate analytical model for dexterous high DoF robots as well as for soft robots requires a lot of engineering knowledge and can be challenging if no accurate parameters are available [12, 16, 39, 70].

The aforementioned advantages of Goal Babbling (e.g., incremental online learning, learning from scratch, learning while behaving, scalability, adaptability, goal-directed learning) make it a promising approach to be implemented for developmental robots as well as for lifelong learning. However, the original Goal Babbling [32] has two main draws back. First, it uses a random goal selection technique to select which goal to learn that remains rather data-hungry and slows down the exploration as well as the learning process, as also shown later in the thesis. Second, Goal Babbling by design learns inverse models, e.g., inverse kinematics, with only one solution for redundant robots which limits the robot's flexibility.

[40] proposed to combine Goal Babbling with an associative dynamic network called "Associative Radial Basis Function network (ARBF)" to enable learning multiple solutions for redundant robots. However, the exploration by Goal Babbling and the solution consolidation in ARBF were done separately. Moreover,

the proposed network works only offline, which is not compatible with lifelong learning.

In this thesis, I follow the idea proposed in [40] for learning multiple solutions with Goal Babbling. I am mainly interested in extending the proposed work on Goal Babbling to fully incremental online learning with simultaneous exploration and solution consolidation. Therefore, the next section briefly discusses the most relevant work related to this thesis without diving deeply into associative networks, which is a rather wide research field and not the main focus of this thesis.

2.6 Associative Dynamic Networks

Different attractor-based dynamical systems have been developed and employed for different learning tasks in robotics, e.g., for imitation learning [100] and motor learning [101–103].

[104] proposed an attractor-based dynamic approach for coupling task and joint space to learn forward and inverse kinematics simultaneously in a single network. The latter work implemented a reservoir computation approach [105] and was demonstrated for smooth task trajectory generation. Learning an inverse model paired with a forward model has been proposed previously as an essential component for motor control and motor learning [91, 92, 106].

[107] extended the idea of implementing attractor-based approaches for learning kinematics to learn and maintain multiple solutions of inverse kinematics. The kinematics redundancy is solved dynamically utilizing multi-stable attractor dynamics similar to [108]. However, the proposed model in [108] has high computational complexity and thus did not demonstrate applicability for real robot applications, in contrast to [40, 41, 107]. [107] combined the ideas of reservoir computing [109] and extreme learning machines (ELMs) [110] to render learning more efficient and alleviate the error of the back-propagation. The trained network is applied in an output feedback loop similar to Jordan-type networks with a feedback loop, e.g., [42, 106, 111, 112].

In [107], the output feedback loop exhibits multi-stable attractor dynamics corresponding to the multiplicity of solutions to the inverse kinematic problem. The estimated output of the network is fed back iteratively to the network until the network settles to one of the attractors depending on its initial state, i.e., the network output will be attracted (converges) to the solution that is closest to its previous state. However, this approach needs to extend the training data set by synthesized sequences to promote attraction to the data samples following the programming dynamics approach proposed in [113]. A similar approach for representing a multi-valued function in a dynamical network with a feedback loop is [114]. However, it integrates additional constraints for the learning objective, which renders training inefficient and requires exhaustive parameter tuning.

[41] proposed an improved variant of the associative dynamic network called "ARBF" to cope with multi-stable dynamics without additional data synthesis, and thus it renders learning more efficient. It utilizes a hidden layer of radial basis functions instead of the sigmoid neurons in [107]. Related to this approach are prototype-based versions of Echo State Network [115] and Recurrent Self-Organizing Maps [116]. However, ARBF uses a recurrent output feedback loop similar to [107] instead of recurrent connectivity in the hidden layer in [115, 116]. ARBF has demonstrated high scalability up to an 11-DoF humanoid [41] and robust performance [117] over several network initializations with high accuracy [40, 41]. It also generalizes well to new situations [40, 41].

There are other common approaches for learning and solving kinematic redundancy, for example, utilizing multiple experts [118] where each solution manifold is modeled separately. However, the number of experts should be selected in advance. It has in general high computational costs, e.g., in exploitation [118] and in training [119]. [120] proposed an online incremental approach with a mixture of linear experts for learning forward and inverse kinematics simultaneously. The number of experts increases incrementally and the experts are allocated automatically. However, selecting a particular solution from a multi-modal distribution remains unsolved similar to other probabilistic approaches for learning kinematics [121]. In contrast, [41, 107] yield directly a proper solution based on the previous robot state, and thus they were employed directly for controlling the robot with smooth motion generation. Other approaches for learning kinematics restrict the inverse models to a single solution [32, 106, 122, 123] which limits the flexibility of the redundant robots.

[40] combined Goal Babbling [32] with the proposed ARBF [41] to learn multiple solutions with Goal Babbling. Goal Babbling has been performed for each desired solution first, and the learned solutions are then consolidated offline utilizing ARBF. This limits the robot adaptability where any change (e.g., in the environment or robot dynamics) requires collecting new data and retraining the system. Besides, the entire data set should be stored which is not possible in lifelong learning. If the full data set is available, radial basis functions can be fitted to the data using k-means -for example- and trained offline [40]. However, in lifelong learning, the full data set is not known in advance. But the model complexity of ARBF must be set beforehand (e.g., the network size), and thus the ARBF is not tailored to the yet unknown learning problem.

In this thesis, I extend the work proposed in [40] and devise an incremental online ARBF (OARBF) to permit online learning of multiple solutions for inverse kinematics with Goal Babbling. The network is constructed incrementally, from scratch, and updated continuously to be tailored to the learning problem.

The incremental learning in lifelong learning has also motivated the idea of incremental recurrent neural networks [77] in order to adapt the network to the learned problem. The proposed networks were implemented for lifelong learning of spatio-temporal representations from videos for continuous object recognition [77]. It is worth mentioning that continuous spatio-temporal patterns can be learned with variational recurrent neural network (RNN) [58, 124]. However, ARBF has the advantage over recurrent and reservoir networks as it alleviates the error of the back propagation and strongly reduces the computational complexity [41, 107]. Hence, ARBF renders learning more efficient which is the main interest and focus of this thesis. Furthermore, efficiency and reduced computational complexity are crucial for direct online training in real robot applications.

How can we devise an efficient online intrinsic motivation method which permits direct online training on physical robots?

3 Interest-Driven Exploration

Developmental robots acquire their sensorimotor skills, e.g., reaching, through lifelong learning in an open-ended environment. On the one hand, random forward exploration of the high-dimensional kinematic space in order to acquire reaching skills, e.g., using motor babbling [93], is infeasible for high DoF robots. On the other hand, trying to reach all foreseen objects in the environment randomly is useless and costly in terms of time, wear, and tear. Hence, these robots should be intrinsically motivated to select what to learn. For example, they should try to select the most informative objects to learn from.

While the previously proposed intrinsic motivation methods are either knowledge-based or competence-based [46, 48], [47] showed experimentally that humans tend to learn by maximizing both their knowledge and their competence. Inspired by this finding, this chapter devises a novel intrinsic motivation method called ”interest measurement” which combines a new competence-based signal called ”forgetting factor” with a new knowledge-based one called ”relative error”. The relative error selects the most informative objects to learn from by shifting the robot's focus toward difficult-to-reach objects. Learning difficult-to-reach goals and generalizing on the simpler ones increases sample-efficiency. Since lifelong learning requires accommodating new knowledge while retaining previously learned experiences [43], the forgetting factor, accordingly, shifts the robot focus toward potentially forgotten, previously learned objects.

The interest measurement method is integrated into a novel hierarchical interest-driven exploration scheme to permit learning robot models online, from scratch, in a learning while behaving fashion and driven by intrinsic motivation. The framework has been evaluated first in an illustrative example with a 10-DoF planar manipulator and compared to the state-of-the-art intrinsic motivation signals. A novel online mental replay method is devised to intensify the robot's experiences in real-world applications. All the proposed methods are demonstrated with a physical 7-DoF Baxter robot arm. The results highlight clearly the advantages gained by the proposed methods: sample-efficiency, robust performance, and fully incremental online learning with instantaneous updates and direct online training on physical robots.

This work has been already published:

- R. Rayyes, H. Donat, and J. Steil, ”Efficient online interest-driven exploration for developmental robots”, IEEE Trans. Cognitive and Developmental Systems, 2020. [125]

- R. Rayyes, H. Donat, and J. Steil, “Hierarchical interest-driven goal babbling for efficient bootstrapping of sensorimotor skills,” in 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020, pp. 1336–1342. [31]

This work has been also presented at the following workshops:

- R. Rayyes and J. Steil, “Hierarchal interest-driven associative goal babbling”, The Annual Conference on Neural Information Processing Systems

(NeurIPS), WiML workshop, Vancouver, Canada, 2019.

- R. Rayyes and J. Steil, "Interest-based exploration with associative goal babbling toward versatile cognitive robots", Robotics Science and Systems (RSS): WiR workshop, Freiburg, Germany, 2019.

- R. Rayyes and J. Steil, "Interest-based exploration with associative goal babbling", The Fourth International Workshop on Intrinsically Motivated Open-ended Learning (IMOL), Frankfurt, Germany, 2019.

3.1 Hierarchical Interest-Driven Exploration Scheme

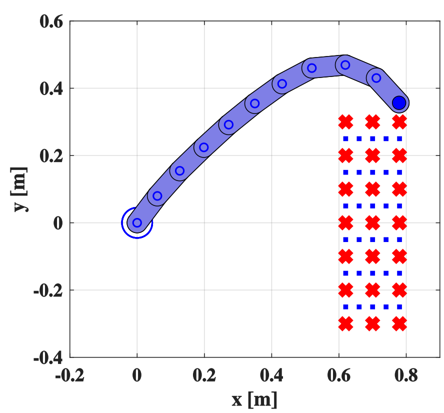


Figure 3.1.1: An illustrative example of the required task with a 10-DoF planar manipulator, each link length is 10 cm. The red crosses and blue dots represent the spatial goals to be reached

This section first describes the task used to demonstrate the proposed methods in this thesis. Fig. 3.1.1 illustrates an example of the required task with a 10-DoF planar manipulator. The robot should learn how to reach some goals in the Cartesian space, which represent the spatial positions of some objects or other physical targets. With some abuse of nomenclature, I sometimes also refer to the goals simply as objects. The learning is online, from scratch, and in a learning while behaving fashion. The robot shall be driven by intrinsic motivation to select what to learn.

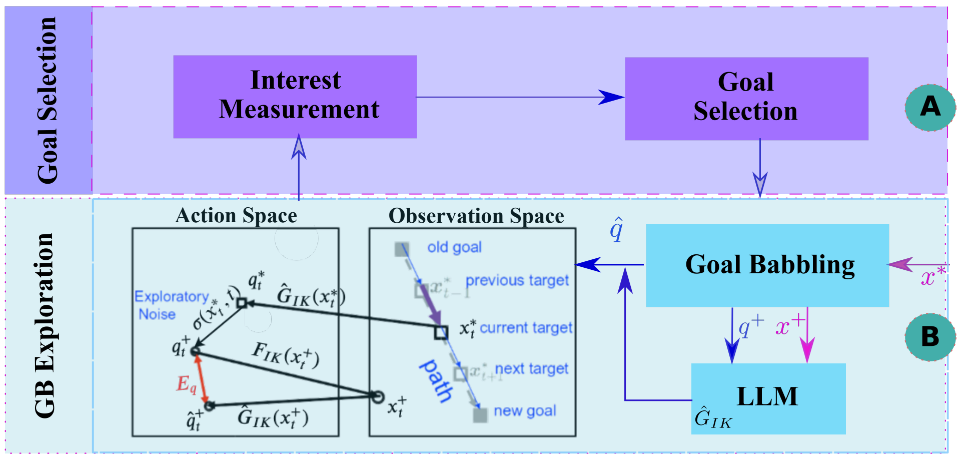


Figure 3.1.2: Hierarchal interest-driven exploration scheme:(A) Goal selection strategy, (B) Goal-directed exploration mechanism

To this aim, a hierarchal interest-driven exploration scheme (cf. Fig. 3.1.2) is developed which consists of a high level of goal selection strategy, and a low level of exploration and incremental approximation of the underlying model:

(A) Goal selection strategy: It utilizes the interest measurement (cf. Sec. 3.2) to guide the robot's exploration driven by intrinsic motivation.

(B) Goal-directed exploration mechanism: It utilizes the interest-driven Goal Babbling (cf. Sec. 3.3) for direct inverse model learning through exploration. This mechanism allows the robot to actively explore its workspace in a learning while behaving fashion. The required model to achieve the task is approximated incrementally and adapted on the fly.

The learning scheme is fully online and updated continuously. Neither storing data sets nor intermediate offline training is required.

3.2 Interest Measurement and Goal Selection

The interest measurement determines which goal the robot will try to attain. At the outset, the robot does not have further knowledge about the goals except that they exist. Therefore, all goals will be interesting to the robot. Once the robot has gathered some knowledge about specific goals, as is measured by the relative error, these become less interesting. Since lifelong learning requires retaining the previously learned experiences [43], the potentially forgotten previously learned goals become again more interesting to the robot by utilizing the forgetting factor. The interest measurement is updated continuously and online to reflect this process and combines knowledge-based (relative error) with competence-based (forgetting factor) signals.

3.2.1 Relative Error

The relative error signal selects the most informative goal to learn from. It compares the performance error on each goal relative to the other ones. The higher the error, the more interesting the goal. A high relative error indicates the goals which haven't been learned yet compared to the other assigned goals. It also indicates the goals which are difficult-to-attain, e.g., goals near the border of the workspace. The relative error is given in Eq. (3.1):

$$\mathrm{RE}(g_i) = \frac{E(g_i) - E_{min}}{E_{max} - E_{min}} \tag{3.1}$$

RE is the relative error of the current goal $g_i \in \mathcal{G}$, $\mathcal{G}$ is the set of the required goals (desired positions) to attain, $E(g_i)$ is the current performance error on the current goal $g_i \in \mathcal{G}$, E_{min} and E_{max} are the current minimum and maximum performance errors over all assigned goals respectively.

On the one hand, the relative error reflects a knowledge-based signal as it measures the instantaneous performance error on each goal, i.e., the difference between the robot prediction and the real goal position. On the other hand, the relative error also reflects a competence-based signal as it measures the relative performance error on the goal set which changes based on the robot's performance progress.

3.2.2 Forgetting Factor

The learned model is approximated incrementally online and updated locally and continuously (cf. Sec. 3.3). Accordingly, the performance of the robot might improve when the robot benefits from previously learned experiences (reaching goals) and generalizes to new ones. The performance might also deteriorate, i.e., the robot forgets potentially about previous experiences. Hence, previously learned goals which start becoming forgotten must become interesting again to the robot. The forgetting factor drives the robot interest toward forgotten previously learned objects (goals) and is given in Eq. (3.2):

$$\left.\begin{aligned} \mathrm{Prog}(g_i) &= \frac{1}{n}\left(\sum_{j=\frac{n}{2}+1}^{j=n} E_j(g_i) - \sum_{j=1}^{j=\frac{n}{2}} E_j(g_i)\right) \\ \mathrm{FF}(g_i) &= \frac{\mathrm{Prog}(g_i) - \mathrm{Prog}_{min}}{\mathrm{Prog}_{max} - \mathrm{Prog}_{min}} \end{aligned}\right\} \tag{3.2}$$

$E_j(g_i)$ is the performance error on the goal $g_i \in \mathcal{G}$, $\mathrm{Prog}(g_i)$ measures the robot performance progress on the goal $g_i \in \mathcal{G}$ over a sliding time window of the last n goal trials (cf. Fig. 3.2.1). $\mathrm{Prog}(g_i)$ increases with the error over the time window. The forgetting factor $\mathrm{FF}(g_i)$ is normalized relative to the minimum Prog_{min} and the maximum Prog_{max} performance progresses on the goals $\mathcal{G}$. The higher the factor, the more interesting the goal. The forgetting factor is a competence-based signal as it measures the robot performance progress, and it is fundamentally similar to the competence-based methods, e.g., the competence measurement [18] and the learning progress signal [19].

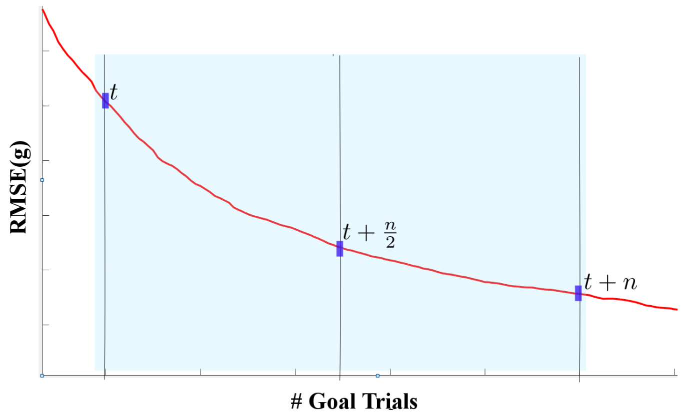


Figure 3.2.1: The robot performance progress on a goal g over n trials. RMSE(g) is the robot performance root mean squared error on the goal g

3.2.3 Interest Measurement

The interest measurement represents an intrinsic motivation learning signal, which is responsible to guide the exploration during the intrinsic motivation learning by shifting the robot's focus toward interesting goals. The interest measurement (also named the interest signal) combines the relative error and the forgetting factor. Accordingly, goals are interesting to the robot either when they are not learned yet or when they are forgotten. The interest measurement is given in Eq. (3.3) :

$$\text{interest}(g_i) = \lambda \text{RE}(g_i) + (1 - \lambda)\text{FF}(g_i) \tag{3.3}$$

where $\lambda \in [0, 1]$ is a weighting factor. Each of interest, RE, FF is normalized in order to have comparable measures on all required goals. Each of these measures is updated on the fly at each time step (on every sample).

3.3 Interest-Driven Goal Babbling

The learning scheme relies on Goal Babbling [32] as an exploration mechanism to permit learning while behaving fashion. Learning the basic reaching skills requires learning the inverse kinematics that assigns to each end-effector position $x \in \mathcal{X} \subset$

Algorithm 1 Interest-Driven Goal Babbling

procedure IGB(x^{home},q^{home}, $\mathcal{X}^*$)
  INITLEARNERS(x^{home}, q^{home})
  for E number of Epochs **do**
    for L/N number of samples **do**
      $(x_t^*) = g_i \in \mathcal{X}^*$ = GOALSELECTION
      for N number of intermediate steps **do**
        generate an intermediate target x_t^*
        estimate the corresponding $\hat{q}_t^*$ for x_t^*
        add exploratory noise σ:
          $q_t^+ = \hat{q}_t^* + \sigma(x_t^*, t)$
        execute q_t^+ and observe (x_t^+)
        compute weight w_t^{gb}
        TRAINLEARNERS(x_t^+, q_t^+, w_t^{gb})
      end for
      INTERESTUPDATE
    end for
    VALIDATIONTEST($\mathcal{G}$)
    INTERESTUPDATE
  end for
end procedure

$\mathbb{R}^n$ the corresponding configuration $q \in \mathcal{Q} \subset \mathbb{R}^m$ that is required to attain it. m is the number of DoF, n is the dimension of the target variable (e.g. $n \in \{2, 3\}$ for the spatial position of the end-effector), $\mathcal{Q}$ is the set of permissible configurations, and $\mathcal{X}$ is the set of the corresponding end-effector positions.

Algorithm. 1 illustrates the Interest-Driven Goal Babbling algorithm. The robot starts exploring from its initial (home) posture q^{home} corresponding to the starting (home) position x^{home}, and tries to reach some goals $g_i \in \mathcal{G}_{train} = \mathcal{X}^*$ which represents some desired positions to be attained. The goals are selected iteratively utilizing the interest measurement (cf. Sec. 3.2). A linear path of N intermediate targets is generated by interpolating between each two selected goals. Note that the term "goal" is used as a predefined desired position, and the term "target" is used as a generated desired position between the goals.

A Local Linear Map (LLM) [35, 44, 126] (cf. Appendix .1) is used as an incremental regression to approximate and update the inverse estimates. Note that any incremental regression technique can be implemented. LLM is chosen because it has already demonstrated a very good accuracy even for estimating complex models (e.g., inverse statics [35]) as well as in real robot applications (e.g., [39]).

The robot tries to reach each generated target and selected goal using the local inverse estimate as follows: A correlated exploratory noise σ is added to the estimated output $\hat{q}_t^*$ ($q_t^+ = \sigma(x,t) + \hat{q}_t^*$) to allow the robot to discover and learn novel outcomes. q_t^+ is executed and the resulting end-effector position x_t^+ is observed. Note that no forward model is required in Goal Babbling, but only visual observation of the effector position.

For redundant robots , it is preferable to learn smooth solutions without switching between configurations. Thus, Goal Babbling tries to select and learn the most efficient solution using the following weighting scheme:

$$\left.\begin{aligned} w_t^{dir} &= \frac{1}{2}(1 + cos\sphericalangle(x_t^* - x_{t-1}^*, x_t^+ - x_{t-1}^+) \\ w_t^{eff} &= \| x_t^+ - x_{t-1}^+ \| \cdot \| q_t^+ - q_{t-1}^+ \|^{-1} \\ w_t^{gb} &= w_t^{dir} \cdot w_t^{eff} \end{aligned}\right\} \tag{3.4}$$

t is the time step, x^* is the desired position, x^+ is the real end-effector position which corresponds to the real configuration q^+, w_t^{dir} assesses whether the actual movement aligns well with the intended one, w_t^{eff} assesses the efficiency of the actual movement, and w_t^{gb} is the sample weight.

(x_t^+, q_t^+, w_t^{gb}) is used to update the local inverse estimate online in a supervised learning fashion in order to minimize the weighted error E_t^q (cf. Fig. 3.1.2, Eq. (3.5)) between the actual q_t^+ and the estimated $\hat{q}_t^+$ configurations as following:

$$\left.\begin{aligned} E_t^q &= w_t^{gb} \| q_t^+ - \hat{q}_t^+ \|^2 \\ \theta_{t+1} &= \theta_t - \eta \cdot \frac{\partial E_t^q}{\partial \theta_t} \end{aligned}\right\} \tag{3.5}$$

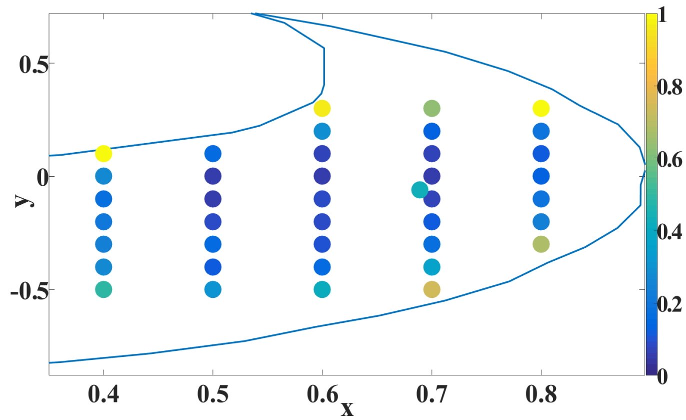


Figure 3.3.1: An illustrative example of the interest measurement

where θ are the LLM parameters (cf. Appendix .1), and η is the learning rate.

q^{home} is used with the weighting scheme for controlling which solution will be learned for a redundant robot. For example, if a 2-DoF planar manipulator starts exploring with an elbow-down home posture, the samples with an elbow-down configuration will receive higher weights than the samples with an elbow-up configuration and vice-versa.

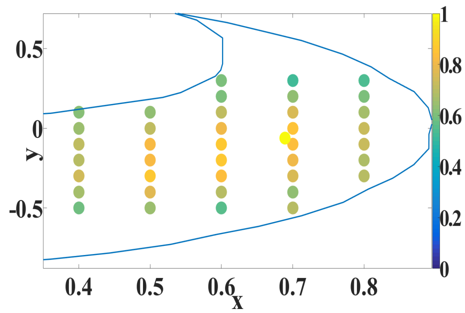
Note that the main differences between the original Goal Babbling [32] and the interest-driven Goal Babbling are:

1. The goals in the original Goal Babbling are selected on a random basis. In contrast, the exploration here is driven by intrinsic motivation and the goals are selected utilizing the interest measurement.

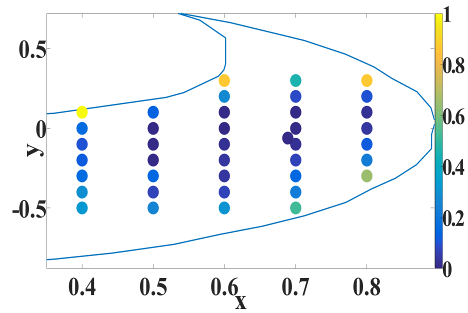
2. x^{home} has been used as a resting position in the original Goal Babbling with a probability $\rho \ll 1$ in order to avoid drifting [32]. In contrast, x^{home} is considered here as one of the predefined goals, and the robot is able to autonomously return to it by using the forgetting factor (cf. Sec. 3.2.2).

Fig. 3.3.1 illustrates an example of the interest measurement after the exploration with a 10-DoF planar manipulator (cf. Fig. 3.1.1). The "yellow" goals indicate the goals which the robot tries to reach most. That means these goals are the most interesting goals for the robot, which are difficult to attain due to the relative error, e.g., goals near the border of the workspace. The goals in dark blue indicate that the robot barely tries to attain them as the learned model benefits from the previous experiences and generalizes well on them. The starting position x^{home} (the green dot in the middle) indicates that the robot returned autonomously to this position due to the forgetting factor.

Fig. 3.3.2 illustrates the effects of the relative error as well as the forgetting factor. The forgetting factor is the highest for the home position (the light yellow dot in the middle) (cf. Fig. 3.3.2(a)), as the robot forgets potentially about initially learned experiences. The relative error makes the robot focuses mostly on the difficult-to-reach goals and implicitly learn the simpler ones (cf. Fig. 3.3.2(b)).



(a) Forgetting Factor

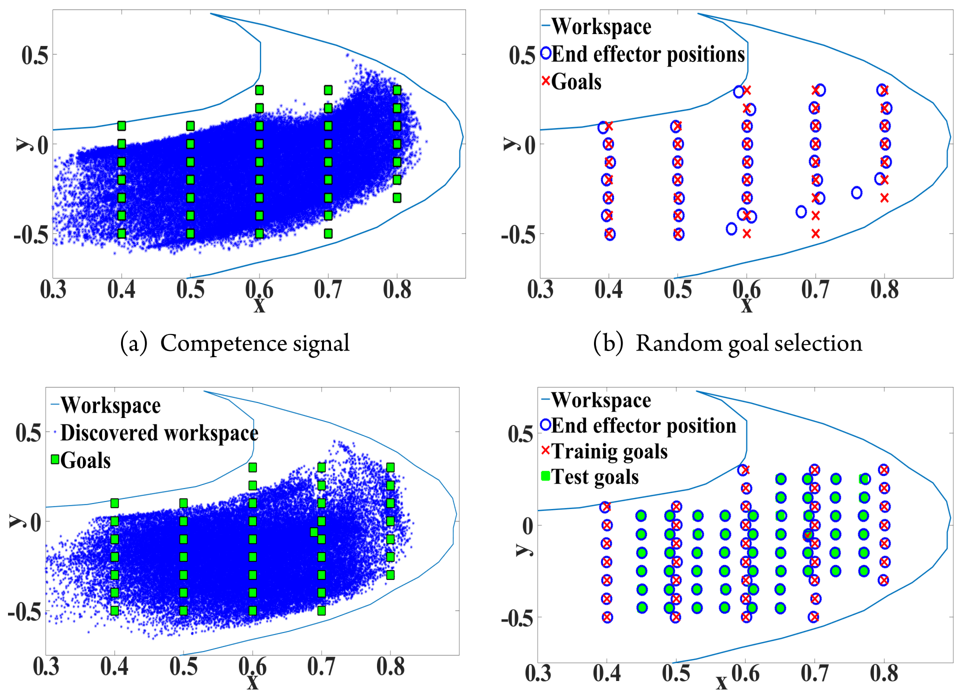


(b) Relative Error

Figure 3.3.2: The effects of the relative error and the forgetting factor

3.4 Comparison with state-of-the-art

In order to demonstrate the advantages gained by the interest measurement, it is compared to other state-of-the-art intrinsic motivation signals. Goal Babbling has been implemented with different learning signals: interest signal (cf. Sec. 3.2), random goal selection signal ([32]), the competence measurement [11, 18], and learning progress [19] in an illustrative setup with a 10-DoF planar manipulator



(a) Competence signal
(b) Random goal selection
(c) Learning progress
(d) Interest signal

Figure 3.4.1: Goal Babbling performance with different learning signals

(cf. Fig. 3.1.1). Note that only competence-based methods have been considered in the comparison, as they have already demonstrated better performance than the knowledge-based ones [47, 49].

The task is to learn how to reach some goals scattered in the workspace illustrated in Fig. 3.3.1. The goals are distributed to include both easy-to-reach goals as well as difficult-to-reach goals near the border of the workspace. Each experiment has been repeated 20 times with 500 epochs, each epoch consists of 100 samples, i.e., 100 time steps. At each epoch, M goals are selected utilizing the corresponding learning signal and the robot tries to reach each of them with 5 intermediate targets (time steps) generated between each two selected goals. Each sample is collected at each time step and represents one observed data point (x_t^+, q_t^+). This sample is generated from the inverse estimate when the robot tries to reach a goal

Table 3.4.1: Interest-Driven Goal Babbling experimental results comparison

Goal Selection	avg. Validation RMSE $[m]$	avg. Test RMSE $[m]$	avg. RMSE std $[m]$
Random selection [32]	$7.4 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$	$5 \cdot 10^{-3}$
Interest measurement	$2.7 \cdot 10^{-3}$	$1.9 \cdot 10^{-3}$	$0.8 \cdot 10^{-3}$
Competence measurement [18]	$5.5 \cdot 10^{-3}$	$2.3 \cdot 10^{-3}$	$1.5 \cdot 10^{-3}$
Learning progress [19]	$9 \cdot 10^{-3}$	$3.4 \cdot 10^{-3}$	$3.8 \cdot 10^{-3}$

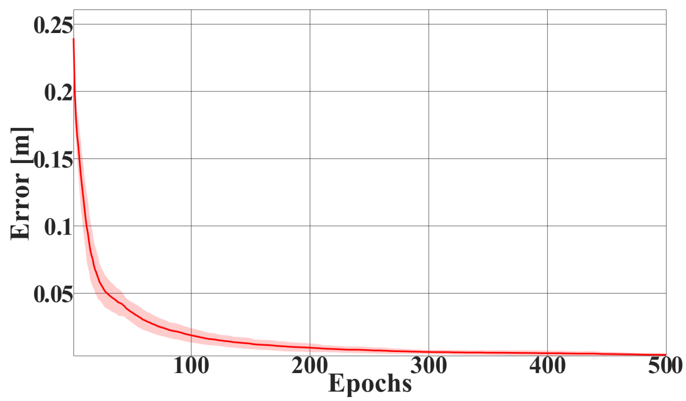


Figure 3.4.2: Interest-Driven Goal Babbling std RMSE for a 10-DoF planar manipulator over 20 experiments

or a generated target (cf. Sec. 3.3).

Table 3.4.1 illustrates the average validation Root Mean Squared Error (RMSE), test RMSE and Standard Deviation (std) of the validation RMSE. The results demonstrate clearly that the interest measurement surpasses other state-of-art learning signals. Over 20 repetitions of each experiment, none of the other learning signals [18, 19, 32] guarantees to achieve all the required goals in each experiment within the defined time frame (500 epochs) as illustrated in Fig. 3.4.1, although the test RMSE for all the implemented learning schemes is almost comparable. In contrast, the interest measurement demonstrated high stability with the minimum std RMSE 0.8 *mm* (the shaded area in Fig. 3.4.2) which surpasses the

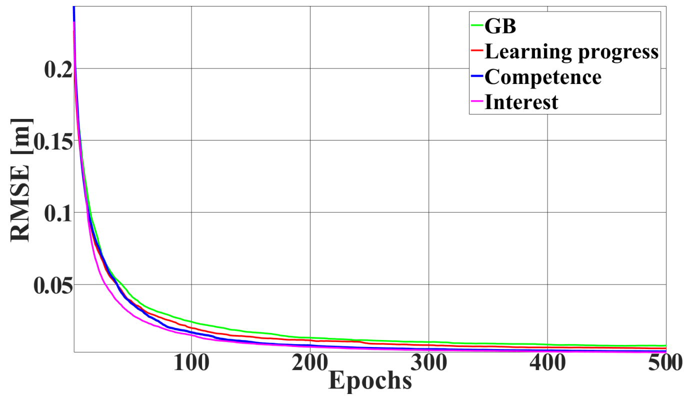


Figure 3.4.3: Mean Performance RMSE of Goal Babbling with different learning signals. GB refers to the original Goal Babbling with random goal selection

other methods and shows a robust performance over all experiments.

The performance error of the interest converges faster than the other signals as illustrated in Fig. 3.4.3 It also achieved the highest precision of the performance accuracy illustrated with 1.9 *mm* validation RMSE and 0.8 *mm* test RMSE.

Note that the test goal set has a different distribution than the validation goal set as illustrated in Fig. 3.4.1(d). The goals, which are located near the border of the workspace, are difficult to reach. This affects the overall validation error in contrast to the test set which is scattered inside the learned workspace. Hence, the validation RMSE is higher than the test RMSE. The validation RMSE has been computed at each epoch to test the robot's performance. As shown in Fig. 3.4.2, the error converges very fast already after 35 epochs.

The good performance of the interest signal is also demonstrated in Fig. 3.4.1(d). All training and test goals are always reached with high accuracy. The red points represent the training goals, the green ones indicate the test goals and the blue circles represent the observed real end-effector positions.

Discussion:

The original Goal Babbling relies on a random selection of the goals [32]. Consequently, all goals receive similar attention from the robot. Whereas the relative error allows the robot to focus on learning the most difficult goals to attain (cf. Sec. 3.2) and generalizes well to the simpler ones. Although the robot tries to attain the difficult goals, it does not run into singularities or undesired configurations. The reason is that the online incremental clustering technique in LLM, which is responsible for adding and initializing new local models, does not change the local behavior of the model and prevents any sudden changes in the estimation [32, 35] (cf. Appendix.1). This is also illustrated in chapter. 4, where the robot extrapolation behavior is reasonable and does not run into unpredictable situations (cf. Fig. 4.4.2), even in a real robot experiment (Fig. 4.5.3). In addition, the generated intermediate targets with Goal Babbling provides smooth continuous motion. Consequently, the solutions unfold gradually from the home position toward more novel outcomes [32].

The original Goal Babbling returns to the home position with a specific probability to avoid drifting, while the interest-driven Goal Babbling could return autonomously to it utilizing the forgetting factor, which allows the robot to focus again on the previously learned forgotten experiences. Forgetting previously learned experiences happened potentially due to the continuous update of the local model (i.e., the performance can both deteriorate or enhance).

In [18], the goals with the high learning progress regardless of whether the robot is learning (RMSE decreases) or forgetting (RMSE increases) receive the highest interest. While the interest signal shifts the robot's focus toward the unlearned goals as well as the forgotten ones. Hence, focusing on difficult-to-reach and generalizing on simpler goals accelerates the learning process. Moreover, the exploration in [18] relies on the combination of a random goal selection and the competence measurement. In contrast, the interest-driven Goal Babbling is completely driven by the interest signal. In addition, the exploration in [11, 13, 18] originally relies on the nearest neighbor method, which is not feasible in online lifelong learning. In contrast, the interest signal is updated online and no data is needed to be stored.

The learning progress [19] gives the highest interest to the goals where the robot learns. This makes the robot focuses only on the simple goals near the home posi-

Table 3.4.2: T-Test for Interest-Driven Goal Babbling vs Goal Babbling with other learning signals

methods	μ_1 $[mm]$	μ_2 $[mm]$	std_1 $[mm]$	std_2 $[mm]$	$d.f$	t-value	p-value	H_0
interest v.s. random	2.7	7.4	0.8	5	38	4.0487	$2.4\cdot10^{-4}$	1
interest v.s. competence	2.7	5.5	0.8	1.5	38	7.3375	$8.6\cdot10^{-9}$	1
interest v.s. learning	2.7	9	0.8	3.8	38	7.6537	$3.2\cdot10^{-9}$	1

tion where the robot can make high progress, without paying much attention to the difficult goals. However, it is worth to mention that despite the slow performance of the learning progress, its main advantage is that it can easily avoid unlearn-able or unreachable goals in contrast to the interest and the competence signals.

T-Test:

An independent-samples t-test [127] is implemented to test whether the difference between the performance accuracies of the interest signal and the other learning signals is significant and not just randomly occurred due to the uncertainties in the data and the learned model.

Table. 3.4.2 shows the results of the t-tests. $H_0 = 1$ indicates the rejection of the null hypothesis at p-value$_{threshold} = 0.05$ significance level. Each of the t-test p-values is less than the significant level, i.e., $t(38) < 0.05$, as illustrated in the table (p-value $<$ p-value$_{threshold}$). This indicates clearly the significant performance accuracy gained by utilizing interest measurement. $d.f = 38$ is the number of DoF in the data population and is given in Eq. (3.6). t-value is given in Eq. (3.7) [127].

$$d.f = n_1 + n_2 - 2 \tag{3.6}$$

$$\text{t-value} = \frac{\mid \mu_1 - \mu_2 \mid}{\sqrt{\frac{\text{std}_1^2 + \text{std}_2^2}{n}}} \tag{3.7}$$

$\mu_1, \ \mu_2$ are the means of the two data populations $\mathcal{D}_1, \ \mathcal{D}_2$ respectively, i.e., the means (averages) of the validation RMSE for the two compared methods over $n = 20$ experiments. $\text{std}_1, \ \text{std}_2$ are the standard deviations of $\mathcal{D}_1, \ \mathcal{D}_2$ respectively, and $n_1, \ n_2$ are the number of data observations for $\mathcal{D}_1, \ \mathcal{D}_2$ respectively. Note that $n_1 = n_2 = n$ denotes the number of the experiments'repetitions.

Which competence-based signal is the best to be utilized in the interest measurement? Learning, Progress, or Forgetting?

The interest measurement combines a knowledge-based signal (relative error) and a competence-based one (forgetting factor) which yields so far the best results in the previous experiments. The forgetting factor, the competence measurement [18], and the learning progress [19] are fundamentally similar. The forgetting factor considers when the robot forgets, the learning progress considers when the robot learns, and the competence measurement focuses on the general progress whether the robot learns or forgets. Hence, this chapter investigates which of these competence-based signal is the best to be combined with the relative error in terms of the performance accuracy and robustness.

Table. 3.4.3 shows clearly that the interest measurement utilizing the forgetting factor (FF) yields the best results overall with the highest accuracy (minimum RMSE) and the highest robustness indicated with the minimum std RMSE of 0.8 *mm*. It also shows clearly that combining any of the competence-based signal with the knowledge-based signal (relative error) enhances significantly the accuracy (RMSE) as well as the robustness of the performance (std) compared to utilizing only the competence or the learning signal (cf. Table. 3.4.1, Table. 3.4.3).

T-Test:

However, as these signals yield comparable accuracies and in order to check whether this small difference between them is significant or not, an independent-samples t-test has been implemented. Table. 3.4.4 illustrates the results of the t-tests. $t(38) > 0.05$ and $H_0 = 0$ indicate that the null hypothesis wasn't

Table 3.4.3: The compassion between the competence-based signals in the interest measurement

The interest measurement	avg. Validation RMSE $[m]$	avg. Test RMSE $[m]$	avg. RMSE std $[m]$
RE and FF	$2.7 \cdot 10^{-3}$	$1.9 \cdot 10^{-3}$	$0.8 \cdot 10^{-3}$
RE and competence measurement [18]	$3.1 \cdot 10^{-3}$	$2 \cdot 10^{-3}$	$1.2 \cdot 10^{-3}$
RE and learning progress [19]	$3.5 \cdot 10^{-3}$	$2.1 \cdot 10^{-3}$	$1.8 \cdot 10^{-3}$

Table 3.4.4: T-Test for interest measurement (FF with RE) vs competence-based signals with RE

Competence-based with RE	avg1 $[mm]$	avg2 $[mm]$	std1 $[mm]$	std2 $[mm]$	d.f	t-value	p-value	H_0
FF v.s. competence [18]	2.7	3.1	0.8	1.2	38	0.9491	0.3486	0
FF v.s. learning progress [19]	2.7	3.5	0.8	1.8	38	1.6947	0.0983	0

rejected at p-value$_{threshold}$ $= 0.05$ significance level (p-value $>$ p-value$_{threshold}$). Accordingly, there is no significant difference between the signals' performances. This indicates clearly that combining the knowledge-based signal (relative error) with any competence-based signal enhances significantly the performance (cf. Table. 3.4.3, Table. 3.4.2).

The next step is to evaluate the proposed learning scheme with the interest measurement on a real robot platform. In order to facilitate deploying the data-driven methods on a real robot and increase further the sample-efficiency, an online mental replay method is first developed in the next section.

3.5 Online Episodic Mental Replay

This section devises a new online replay method called "Online Episodic Mental Replay (OEMR)". OEMR intensifies the robot's experiences in real-world applications. It allows rapid online updating of the incrementally learned model using the last training epoch only.

During each epoch, M goals are selected and the robot tries to reach them. N intermediate targets are generated between each two selected goals. Hence, each epoch consists of $M \times N$ samples which are stored temporarily in the replay buffer. At the end of each epoch, these samples are replayed imaginary to update the learner.

During the learning, the model is approximated incrementally and updated continuously, where only one gradient descent step is performed for each received sample (cf. Eq. (3.5)). Therefore, performing OEMR accelerates the convergence of the learner which reduces rapidly the required samples to approximate the model.

In contrast to the previously proposed replay methods, OEMR neither requires storing the full data set nor to augment the data set.

3.6 Interest-Driven Exploration with a Physical 7-DoF Baxter

The Interest-Driven Goal Babbling with OEMR has been implemented on the physical 7-DoF left arm of Baxter robot (cf. Fig. 3.6.1) in order to demonstrate the applicability as well as the gained efficiency by the proposed methods. A supplementary video showing the experiment is available at `https://youtu.be/W6tB-7fos4A` [128].

Baxter has a positioning accuracy of 5 *mm* [45], and learning while behaving in practice produces highly noisy data. Therefore, the task is very challenging on this robot. The Baxter sampling rate using MoveIt - Motion Planning Framework [129] is 3 *sec* for each time step. For example, if 7 intermediate targets are generated between every two goals, at least 3×7 sec is needed to reach the goal and collect these samples. The parameter set is $\{\eta = 0.0725, \sigma = 0.0452, r = 0.0869, \lambda = 0.5\}$, where η is the learning rate, σ is the exploratory noise (cf. Sec. 3.3), r is the

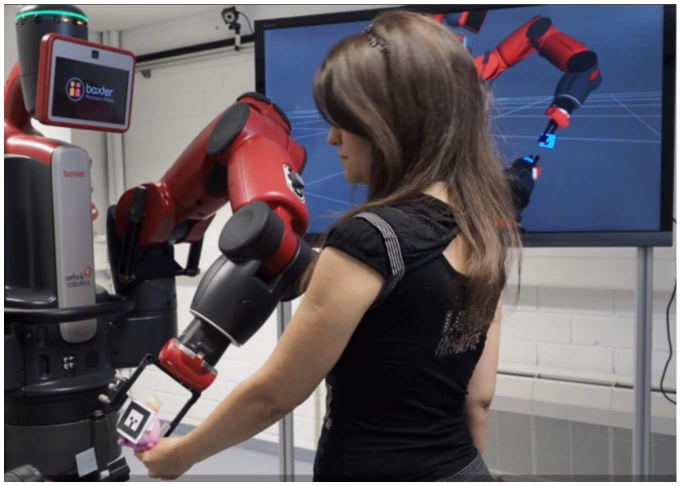


Figure 3.6.1: Baxter reaching a detected object after the exploration phase

radius parameter of LLM [32, 35, 44] (cf. Appendix .1), and λ is the weighting factor of the interest signal (cf. Sec. 3.2).

The task is here, how the robot can learn to reach some objects (desired spatial positions) as shown in Fig. 3.6.1 without any prior-knowledge of the model, online, from scratch, in a learning while behaving fashion, and driven by its own interest.

As shown in the video, the experimenter first shows the robot a desired workspace to be explored. A three-dimensional virtual goal grid is created inside the defined volume and illustrated in a 3D visualizer (rviz provided by ROS) as shown in Fig. 3.6.2. The goals are scattered in a cuboid shape, with a vertical and horizontal distance of 10 cm between every two adjacent goals. Note that these virtual goals are used for the exploration in the real experiment with the physical robot (cf. Fig. 3.6.3).

The robot starts exploring the determined workspace from its starting position (home position) trying to reach these virtual goals as illustrated in Fig. 3.6.3 and in the video. The goals are selected iteratively utilizing the interest measurement.

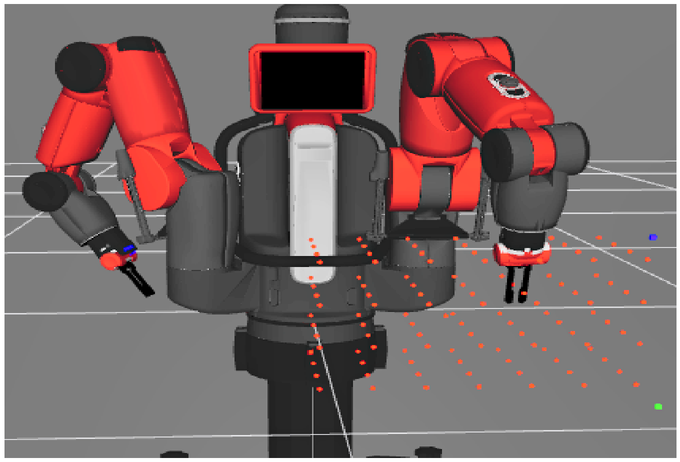


Figure 3.6.2: Virtual goal grid visualized in rviz

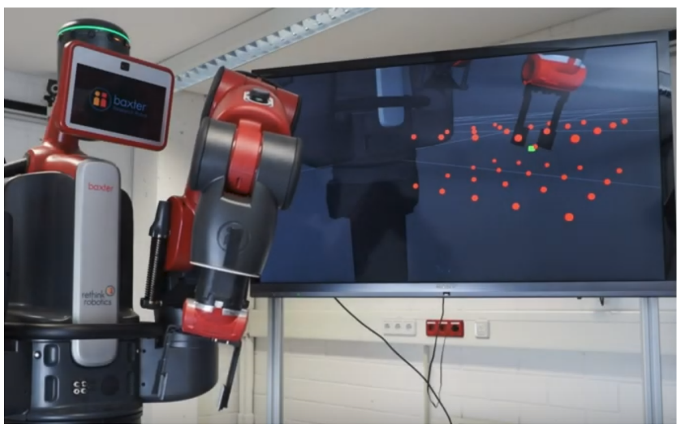


Figure 3.6.3: Baxter explores the detected workspace driven by the interest measurement and utilizing the virtual goals

Each goal trial consists of N intermediate targets, which varies depending on the distance between the selected goals. After each epoch (1000 samples, i.e., 1000 time steps), the robot performance is evaluated on the virtual goal set and OEMR is performed.

Only 4 epochs were needed to learn the detected workspace and reach 41 goals with 6.7 *mm* validation RMSE within 3 *hours* and 20 *min* of direct online training on the physical Baxter robot, with sampling rate of 3 *sec* per sample (i.e., per time

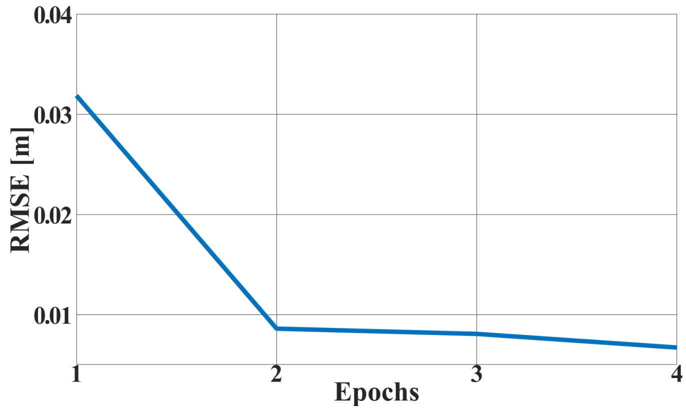


Figure 3.6.4: The performance RMSE of Baxter

step). Fig. 3.6.4 shows the validation RMSE, the error converged very fast already after the first training epoch. The other three epochs were performed to enhance the performance accuracy and to demonstrate the stability of the learning system. The robot demonstrated very robust performance despite the instantaneous update of the interest measurement.

In the test phase, the robot was able to reach 93 new virtual goals randomly scattered in the explored workspace with a test RMSE of 7.8 *mm* which is acceptable compared to the low positional accuracy of the Baxter of 5 *mm*. The robot performance is also tested on a real detected object. The robot could reach the object in different positions as illustrated in Fig. 3.6.1 and in the video.

The experiment demonstrated clearly the applicability of the proposed methods and the high efficiency gained, where only 4 epochs were required to approximate the internal model with reasonable accuracy within a few hours of direct online training on a real robot.

3.7 Conclusion

This chapter devised a novel intrinsic motivation method called "interest measurement". This method combines knowledge-based with competence-based signals.

The proposed knowledge-based signal called ”relative error” selects the most informative goals to learn from by shifting the focus of the robot toward difficult-to-attain goals. The proposed competence-based signal called ”forgetting factor” assures the lifelong learning concept by shifting the robot's interest toward potentially forgotten learned goals. The interest measurement method is within the few intrinsic motivation methods which have been demonstrated in real robot applications.

The interest measurement method was illustrated within a hierarchical interest-driven exploration scheme to acquire basic reaching skills in a learning while behaving fashion. The interest measurement method outperformed other state-of-the-art methods in terms of accuracy and robust performance. The results also demonstrated the advantage of the relative error signal, where combining it with any competence-based signal (learning, forgetting, progress) improved significantly the performance of these signals.

The chapter also proposed a new online mental replay method called ”online episodic mental replay”, which is a memory of only the most recent learning epoch. It facilitates deploying data-driven learning methods on physical robots by intensifying the robot's experiences. The main advantage of this replay method is that it requires neither storing the full data set which is not feasible in lifelong learning nor augmenting the sample space which might not be known in advance, in contrast to the other state-of-the-art replay methods.

All the proposed methods in this chapter were demonstrated with direct online training on a physical 7-DoF Baxter robot arm. The results highlighted the efficiency gained by the proposed methods in terms of the number of required samples to learn the model with reasonable accuracy. First, the relative error signal searches for the most informative goals to learn from. This increases the efficiency by reducing the number of training samples. Second, the online episodic mental replay updates the model locally and rapidly. This helps local models to coverage fast and consequently require less samples to update. The results showed clearly that the robot was able to learn a full goal set by learning only some goals and generalizing on the others. The robot was able to learn the required task with less than 4 hours of direct online training. Only 4000 training samples were required to learn the model with reasonable accuracy compared to the Baxter positioning accuracy.

The proposed learning scheme is fully online, each received sample is processed immediately and the learner is updated on the fly continuously without any offline or batch learning. Despite the instantaneous continuous update of the interest measurement, the robot demonstrated a robust performance and a consistent interest over 20 experiments. Increasing efficiency and providing a possible instantaneous online update facilitated the real robot application. The learning scheme also demonstrated high stability even in the presence of noisy data, the instantaneous online update of the learner and the interest signal, and the low accuracy of Baxter.

How can we integrate intrinsic motivation with learning from observation and enable the robot to autonomously decide when and how to explore?

4

Extrinsic-Intrinsic Motivation Learning

In the previous chapter, a novel intrinsic motivation method named "interest measurement" was developed to guide autonomously the robot's exploration driven by its interest. This method provides an efficient goal selection strategy to enable the robot to efficiently select what to learn. Since developmental robots also share the environment with humans, detecting humans' goals can point out important outcomes to learn as well as give important cues to where to explore.

This chapter integrates the interest measurement with a novel observational learning method in a new extrinsic-intrinsic motivation learning scheme. The robot's exploration, which is autonomously guided by intrinsic motivation, is additionally guided by observing human demonstrations' outcomes. To this aim, this chapter adapts and adopts the learning from observation concept as an extrinsic motivation method to enable the robot to benefit from humans. Extrinsic motivation in this thesis means that, when the robot observes human demonstrations, it is motivated to learn to achieve similar observed outcomes by exploiting and adapting its internal model. Note that human demonstrations in this thesis mean that, the human reaches objects while the robot is observing his outcomes.

This chapter first explains the general architecture of the extrinsic-intrinsic motivation learning scheme, and then focuses on devising the new methods required for extrinsic motivation learning:

- Novelty detection method, based on descriptive statistics, to check if the newly detected task (goal) from observing a human demonstration is novel.

- Novelty degree, based on descriptive statistics, to measure how informative the detected novel task (goal) is for the robot.

- Probabilistic extrinsic signal to guide the robot's exploration during extrinsic motivation learning. It allows the robot to expand its knowledge about the workspace and learn new tasks (goals) observed from human demonstrations. It increases sample-efficiency by trying to select the most informative goals to learn from.

The novelty detection and novelty degree methods permit the robot to decide on its own whether to explore further to gain more knowledge, or to exploit its internal learned model to achieve similar outcomes as humans. These methods also enable the robot to decide which exploration strategy it should follow: self-exploration driven by intrinsic motivation or learning by observing human demonstrations. In addition, this chapter develops further the interest signal into a probabilistic intrinsic signal to provide a fully probabilistic goal selection strategy utilizing the intrinsic and extrinsic signals. This advancement can cope better with situations where there are only a few goals to learn, as it avoids getting stuck in difficult-to-learn goals. These methods together provide a novel and unique integrative scheme to guide the robot's exploration, which in principle could be utilized in different learning scenarios.

The framework has been evaluated first in an illustrative robot setup with a 10-DoF planar manipulator and then on a physical 7-DoF Baxter robot arm. The results highlight clearly the advantages gained by the proposed learning scheme which leverages intrinsic and extrinsic motivation learning benefits: fast learning and adaptation, expanding exploration toward novel areas of the workspace benefiting from observing human demonstrations, robust exploration strategy, managing autonomously the exploration-exploitation trade-off, sample-efficiency, and fully incremental online data-driven learning with direct online training on physical robots.

The main work presented in this chapter has been developed during my research internship at Sony Computer Science Laboratories (CSL) in Tokyo - Japan, Oct 2019 - Mar 2020, working with Dr. Michael Spranger. The Baxter experiment has been conducted in cooperation with Heiko Donat at IRP, TU Braunschweig.

This work has been already accepted for publication:

- R. Rayyes, H. Donat, J. Steil, and M. Spranger, "Interest-driven exploration with observational learning for developmental robots," IEEE Trans. Cognitive and Developmental Systems, in press. [130]

4.1 Online Extrinsic-Intrinsic Motivation Learning Scheme

This section briefly describes the overall task used to demonstrate the new methods. The task here is to learn how to reach the spatial positions of some objects (goals $\mathcal{G}$) in the Cartesian space online, from scratch, and in a learning while behaving fashion. The robot shall be driven by intrinsic motivation and be able to benefit also from observed human demonstrations' outcome to accelerate learning. It shall decide on its own when to explore and which exploration strategy to follow.

The robot needs first to acquire some knowledge about its internal model. For this aim, the robot tries to reach some goals $\mathcal{G}_{train}$ driven by intrinsic motivation (e.g., utilizing the probabilistic intrinsic signal (cf. Sec. 4.3.2) or the interest sig-

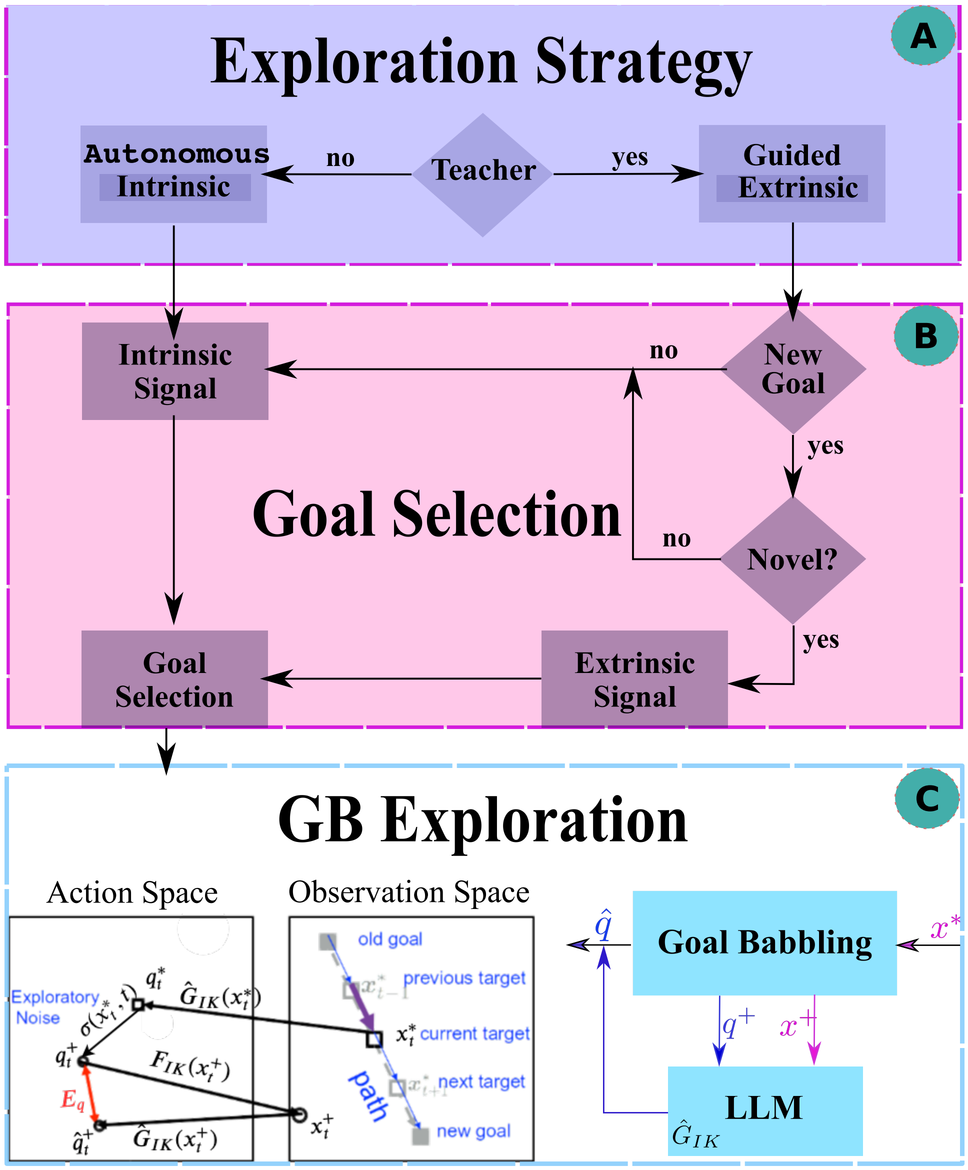


Figure 4.1.1: Extrinsic-intrinsic motivation learning scheme: (A) Exploration strategy, (B) Goal selection strategy, (C) Goal-directed exploration mechanism

nal (cf. Sec. 3.2)). When a human demonstration is observed and a new goal is detected, the novelty of the detected goal is measured by how much the robot has knowledge about it and how much this goal is informative to the robot. If novel goals are detected, the robot is motivated to learn them and achieve similar outcomes as humans by exploiting and adapting its internal model and expanding its knowledge about the environment.

It is important to differentiate between two terms: "new" and "novel". The robot could have enough knowledge to achieve a new goal. However, if a new goal is a novel goal, "novel" indicates that the robot does not have enough knowledge to achieve this goal and it needs to explore and learn more. In this thesis, I assume that human demonstrations only occur when it is necessary to point out important goals to learn.

Fig. 4.1.1 illustrates the extrinsic-intrinsic motivation learning scheme. The learning scheme comprises two high levels of exploration strategy and goal selection strategy, and a low level of exploration mechanism and incremental approximation of the underlying model:

(A) Exploration strategy: This strategy determines whether the robot's exploration is guided by intrinsic or extrinsic motivation.

(B) Goal selection strategy: When novel goals are detected from observing human demonstrations, the goals are selected utilizing the extrinsic signal (cf. Sec. 4.3.1). Otherwise, the goals are selected utilizing the intrinsic signal (cf. Sec. 4.3.2).

(C) Goal-directed exploration mechanism: The underlying exploration mechanism relies on Interest-Driven Goal Babbling (cf. Sec. 3.3) to reduce the search space and leverage the advantages of Goal Babbling.

The learning scheme is updated online on the fly. Neither storing data sets nor intermediate offline training is required.

4.2 Novel Goal Detection

Assume that the robot can infer new goals $\mathcal{G}_{new}$ to learn by observing human demonstrations. When a new goal is induced, it is important first to detect whether this goal is novel or not, and how informative this goal is for the robot. To this aim, two methods have been devised in this section: (i) *Novelty Detection* for new goals and (ii) *Novelty Degree* to measure the goals' novelty, which reflects how much knowledge the robot has about them. Based on this information, the robot can take an informed decision on whether it needs to explore further to achieve the new goals, or it can simply exploit its internal model.

These methods are devised based on descriptive statistics as explained in the following sections.

4.2.1 Descriptive Statistical Method Analysis "Five-Number Summary"

The novelty here is devised based on descriptive statistics of the distribution D of performance errors $E(g_i)$ on the trained goals $g_i \in \mathcal{G}_{train}$. It uses the "five-number summary" [131] which relies on the interquartile range (IQR) to be robust against outliers. It is also independent of the data distribution. The data distribution is described with the "five-number summary" (Median, Max, Min, Q_1, Q_3) as follows:

- Median is the middle of the data set, i.e., the data value at 50% of D.
- First quartile (the lower quartile) Q_1 is the middle number between the smallest value and the median, i.e., the data value at 25% of D.
- Third quartile (the upper quartile) Q_3 is the middle number between the largest value and the median, i.e., the data value at 75% of D.
- Interquartile range IQR is given in Eq. (4.1). IQR represents the variability of the data based on dividing the data set into quartiles. IQR is more resistant to the outliers than the variance (σ^2) and the standard deviation (std).

$$\mathrm{IQR} = Q_3 - Q_1 \tag{4.1}$$

- Maximum excluding any outlier, which is also called the upper inner fence (UIF), is given in Eq. (4.2):

$$\text{Max} = \text{UIF} = Q_3 + 1.5 \times \text{IQR} \tag{4.2}$$

- Minimum excluding any outlier, which is also called the lower inner fence (LIF), is given in Eq. (4.3):

$$\text{Min} = \text{LIF} = Q_1 - 1.5 \times \text{IQR} \tag{4.3}$$

Any data point with a value larger than Max or smaller than Min is detected as an outlier.

4.2.2 Novelty Detection

This method detects whether the new goals induced from observing human demonstrations are novel or not. It also detects the goals which have not been learned well during the intrinsic motivation learning. To do so, a novelty detection threshold should be inferred automatically based on the knowledge that the robot has. The acquired knowledge during the intrinsic motivation learning, which is measured as the performance error on $\mathcal{G}_{train}$, could be used as a threshold for detecting novelty. However, not all goals in $\mathcal{G}_{train}$ might be learned well. Hence, an error threshold for detecting novelty E_{thr} is defined as the Max value, i.e., the upper inner fence (UIF) value (cf. Eq. (4.4), Eq. (4.2)):

$$E_{thr} = E_{Q_3} + 1.5 \times E_{IQR} \tag{4.4}$$

where

$$E_{\text{IQR}} = E_{Q_3} - E_{Q_1} \tag{4.5}$$

E_{Q_3} is the third quartile of the performance error distribution on $\mathcal{G}_{train}$, E_{Q_1} is the first quartile of the error distribution, and E_{IQR} is the variance of the performance errors on $\mathcal{G}_{train}$ (cf. Eq. (4.1)). The novelty detection $\mathcal{N}$ is given accordingly as

$$\mathcal{N}(g_i) = \begin{cases} 1 & E(g_i) > E_{thr} \\ 0 & otherwise \end{cases} \tag{4.6}$$

$E(g_i)$ is the performance error on the goal $g_i \in \mathcal{G}$.

4.2.3 Novelty Degree

The novelty degree $\mathcal{ND}$ measures how much the goal is novel and informative to the robot. $\mathcal{ND}$ of a novel goal g_{novel} is the difference between the test performance error $E(g_{novel})$ on this novel goal and the error threshold E_{thr} (cf. Eq.(4.4)) relative to the variability of the performance errors E_{IQR} on the goals (cf. Eq. (4.5)).

$$\mathcal{ND}(g_{novel}) = \frac{E(g_{novel}) - E_{thr}}{E_{IQR}} \tag{4.7}$$

Normalized Novelty Degree

The normalized novelty degree $\mathcal{NND}$ is given in Eq. (4.8).

$$\mathcal{NND}(g_i) = \max\left(\frac{E(g_i) - E_{thr}}{E_{max} - E_{thr}}, 0\right) \tag{4.8}$$

$\mathcal{NND} \in [0, 1]$, and E_{max} is the maximum performance error over all goal set $\mathcal{G}$. The novelty degree is normalized relative to the E_{thr}. Note that all goals on which the performance error is less than E_{thr} are not novel and their corresponding $\mathcal{NND}$ and $\mathcal{N}$ are 0.

If a novel goal is detected from an observed human demonstration's outcome, the exploration strategy in the learning scheme (c.f. Fig. 4.1.1) switches automatically to the extrinsic motivation learning. Otherwise, the robot can achieve the new goals by exploiting its internal model.

The exploration in the extrinsic motivation learning is guided utilizing the probabilistic extrinsic signal which is established in the next section.

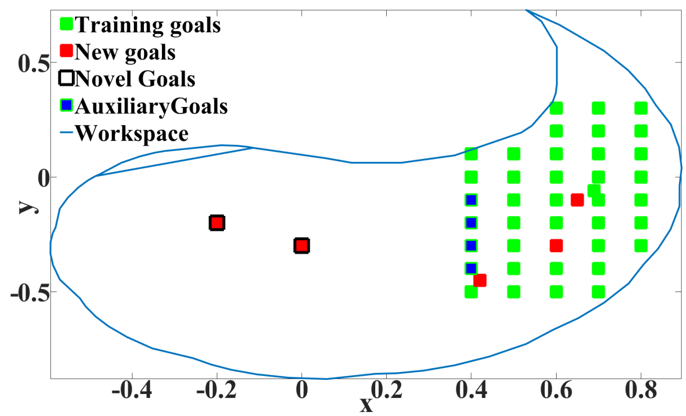


Figure 4.3.1: An illustrative example of the goal sets in the workspace of a 10-DoF planar manipulator

4.3 Probabilistic Goal Selection Strategy

4.3.1 Probabilistic Extrinsic Signal

Novel goals introduce a new challenge to the robot to handle them properly because they might require to extend learning to new areas of the workspace, where there are no known goals to be found. Therefore, it is good to expand the exploration to learn more novel outcomes between the previously discovered workspace and the newly detected goals, in order to avoid having any gab in the robot's knowledge. To define a suitable strategy to handle this situation, different goal sets have to be distinguished first. Fig. 4.3.1 represents an illustrative example of five different goal sets scattered in the workspace of a 10-DoF planar manipulator with specific joint limits. The green dots represent the training goal set $\mathcal{G}_{train}$, i.e., the goal set which is used in the intrinsic motivation learning. The blue ones represent the auxiliary goal set $\mathcal{G}_a \subseteq \mathcal{G}_{train}$ which is explained later in this section. The red dots illustrate the newly detected goals $\mathcal{G}_{new}$ from observing human demonstrations, and the black squares indicate the novel goals $\mathcal{G}_{novel}$ which are detected by the novelty detection $\mathcal{N}$ (cf. Eq. (4.6)). Accordingly, the full goal set is $\mathcal{G} = \mathcal{G}_{new} \cup \mathcal{G}_{train}$.

Remember that the novelty degree method detects novel goals from the full $\mathcal{G}$. Each goal with a higher performance error than E_{thr} is marked as a novel goal. Accordingly, the novel goals are not only goals from the new goal set but also the goals which have not been learned well during the intrinsic motivation learning (cf. Sec. 4.2.2). Therefore, the novel goal set is $\mathcal{G}_{novel} \subset \{\mathcal{G}_{new} \cup \mathcal{G}_{train}\}$.

Now assume that the training set is learned well through the intrinsic motivation learning. Accordingly, $\mathcal{NND}$ (cf. Eq. (4.8)) is (0) for each goal $g_i \in \mathcal{G}_{train}$. For example, in the experiment illustrated in Fig. 4.3.1, five new goals are detected from observing human demonstrations, and only two, which are located outside the learned workspace, are detected as novel goals utilizing the novelty detection $\mathcal{N}$ (cf. Eq. (4.6)). Their corresponding novelty degree measures are: $\mathcal{ND} = \{83.2655, 118.1042\}$ *IQR*, $\mathcal{NND} = \{0.6794, 1\}$. Accordingly, it is not enough to rely only on $\mathcal{NND}$ to drive the exploration, as this will shift the robot's focus toward these two novel goals only. It is better to expand the exploration to learn additional novel outcomes between the previously discovered workspace and the newly detected goals, as discussed before. To this aim, a probabilistic extrinsic signal is devised to drive learning by choosing goals. However, it is futile to iterate between already well-learned goals. In order to devise an efficient probabilistic goal selection strategy, an auxiliary goal set is defined first.

The Auxiliary Goal Set:

The auxiliary goal set $(\mathcal{G}_a \subset \mathcal{G}_{train})$ is defined by determining k nearest neighbors from $\mathcal{G}_{train}$ for each goal $g_{novel} \in \mathcal{G}_{novel} \cap \mathcal{G}_{new}$

$$k(g_{novel}) = 2\ \max\left(\text{round}\left(\frac{(\min(\text{dist}(g_{novel}, g_{train}))}{r}\right), 1\right) \tag{4.9}$$

$k(g_{novel})$ is the number of the auxiliary neighbored goals for each $g_{novel} \in \mathcal{G}_{novel} \cap \mathcal{G}_{new}$, dist is the distance measure, e.g., Euclidean distance, between $g_{train} \in \mathcal{G}_{train}$ and $g_{novel} \in \mathcal{G}_{novel} \cap \mathcal{G}_{new}$, and r is the maximum distance between two neighbored goals in $\mathcal{G}_{train}$. Hence, at least 2 auxiliary goals are defined for each novel new goal. The number of the auxiliary goals increases relative to the distance between the

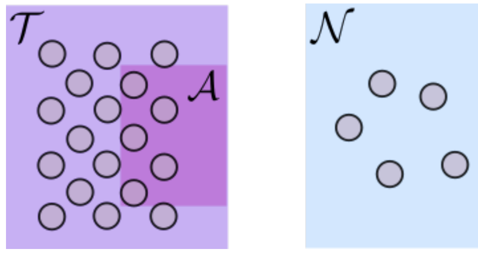


Figure 4.3.2: The goal sets in the extrinsic motivation learning

novel new goals and the training set. This is important to expand the exploration to a wider range of the workspace.

Note that a novel goal might be a not so well learned training goal or a newly induced goal ($g_{novel} \in \mathcal{G}_{novel} \subset \{\mathcal{G}_{new} \cup \mathcal{G}_{train}\}$). However, the auxiliary goals are only defined for the novel newly induced goals ($g_{novel} \in \mathcal{G}_{novel} \cap \mathcal{G}_{new}$).

Probabilistic Extrinsic Signal

A probabilistic extrinsic signal is proposed as a probabilistic goal selection strategy to guide the exploration during the extrinsic motivation learning. This approach is a discrete weighted random number sampling, and the probability mass function (PMF) for the goal selection is given in Eq. (4.10):

$$P(g_i) = \begin{cases} \rho_0 & \text{if } g_i \in \mathcal{G} \setminus \{\mathcal{G}_{novel} \cup \mathcal{G}_a\} \\ \rho_a & \text{if } g_i \in \mathcal{G}_a \\ \rho_n & \text{if } g_i \in \mathcal{G}_{novel} \end{cases} \tag{4.10}$$

$P(g_i)$ is the probability of selecting goal $g_i \in \mathcal{G}$. ρ_0, ρ_a, ρ_n are goal selection probabilities with $\rho_0 > 0$, $\rho_a > \rho_0$, and $\rho_n = f(\mathcal{NND}(g_i)) \in (\rho_0\ 1]$ ($\mathcal{NND}$ cf. Eq. (4.8)), Accordingly, the goal sets (cf. Fig. 4.3.1) can be divided into three main goal sets for the extrinsic motivation learning as illustrated in Fig. 4.3.2.

- $\mathcal{T} = \mathcal{G} \setminus \mathcal{G}_{novel}$ with cardinality $n_{\mathcal{T}} =| \mathcal{T} |$. It represents the goals that are already learned well.

- $\mathcal{A} = \mathcal{G}_a \subseteq \mathcal{G}_{train}$ with cardinality $n_{\mathcal{A}} =\mid \mathcal{A} \mid$. It represents the auxiliary goals, which are also learned well.

- $\mathcal{N} = \mathcal{G}_{novel}$, with cardinality $n_{\mathcal{N}} =\mid \mathcal{N} \mid$. It represents the detected novel goals.

The selection set $\mathcal{S}$; accordingly, is defined as follows:

$$\mathcal{S} = \{t, a, n\} \tag{4.11}$$

a represents selecting a random goal $g_i \in \mathcal{A}$. Similarly, if t is selected, one goal $g_i \in \mathcal{T}$ will be selected randomly. If n is selected, a novel goal from $\mathcal{N}$ will be selected with a probability based on its novelty degree.

The weighting scheme for the item selection in $\mathcal{S}$ is defined based on $\mathcal{NND}$ (cf. Eq. (4.8)) and satisfies Eq. (4.10) as follows:

$$\begin{cases} w(t) = w_0 > 0 \\ w(a) = w_a > w_0 \\ w(n) = \sum_{g_i \in \mathcal{N}} w_n(g_i) : w_n(g_i) = \lceil (\mathcal{NND}(g_i) \times 10 \rceil + w_0 \end{cases} \tag{4.12}$$

Note that sampling from $\mathcal{S}$ is an ordered sampling with replacement. It will be inefficient to iterate between the well-learned auxiliary goals with $\rho_a \gg 0$ (cf. Eq. (4.10)). In order to accelerate learning, $g_i \in \mathcal{G}_a$ is selected with a probability p_a only if the previously selected goal is a novel goal, i.e., $g_{i-1} \in \mathcal{G}_{novel} \equiv \mathcal{N}$.

Accordingly, the extrinsic probabilistic signal guides the exploration during the extrinsic motivation learning utilizing this weighting scheme for goal selections. The probabilities for selecting each action for the set $\mathcal{S}$ as well as for selecting each goal from the goal sets are calculated in Appendix .2.

4.3.2 Probabilistic Intrinsic Signal

In order to devise a fully probabilistic goal selection strategy in this learning scheme, this chapter develops further the interest measurement [31] as a probabilistic intrinsic signal instead of deterministic one. The probabilistic intrinsic signal can cope better in the situations where there are only a few goals as it avoids getting stuck in difficult-to-learn goals (see Appendix .3 for the evaluation).

Each goal $g_i \in \mathcal{G}_{train}$ is weighted with w_{int} (cf. Eq. (4.14)) based on its interest measurement (cf. Sec. 3.2), and it is selected with a probability $P(g_i)$ which is determined by the probability mass function (PMF) given by Eq. (4.16).

$$f_a(g_i) = \exp(\text{interest}(g_i) + a) \tag{4.13}$$

$$w_{int}(g_i) = f_a(g_i) - \text{fmin}_a + 1 \tag{4.14}$$

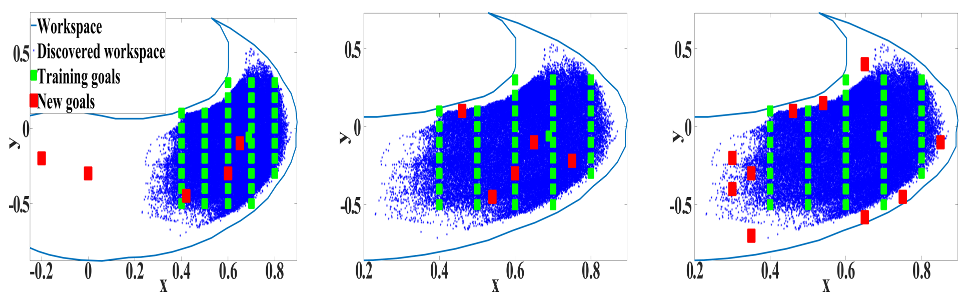
$$\text{fmin}_a = \min\{f_a(g_j) : \forall g_j \in \mathcal{G}_{train}\} \tag{4.15}$$

$$P(g_i) = \frac{w_{int}(g_i)}{\sum_{j=1}^{N} w_{int}(g_j)} \tag{4.16}$$

a is the importance factor which scales the interest measurement exponentially. If a is too small, all goals will have almost equal probabilities to be selected. If a is too large, it will shift the focus of the robot toward the goals with the highest interest only. fmin_a is the minimum value of the function f_a and is corresponding to the minimum interest value. Adding 1 in Eq. (4.15) serves to avoid assigning (0) probability to any goal. N is the number of goals. Note that this goal selection approach is a discrete weighted random number sampling, employing ordered sampling with replacement. The intrinsic and the interest signals are updated online at each time step on every sample.

In the beginning, the interest measurement will assign 1 to all goals, as the robot does not have any prior-knowledge about them. Hence, all goals are selected with equal probability $P(g_1) = P(g_2) = \ldots = P(g_n) = \frac{1}{N}$. When the robot starts exploring and gains some experiences (x, q), where x is the observed end-effector position and q is the corresponding configuration, the probability for each goal to be selected is conditioned with the incremental knowledge the robot gains: $P(g_i|x_1, q_1), P(g_i|x_1, q_1, x_2, q_2), P(g_i|x_1, q_1, x_2, q_2, \ldots, x_m, q_m, \ldots)$. This is expressed implicitly in the probabilistic intrinsic signal as the interest measurement is updated continuously to reflect the robot's knowledge about the goals.

4.4 Extrinsic Motivation Learning Evaluation



(a) case (1): $\mathcal{G}_{new}$ scattered inside and outside the discovered workspace

(b) case (2): $\mathcal{G}_{new}$ scattered inside the learned workspace

(c) case (3): $\mathcal{G}_{new}$ scattered near the border of the discovered workspace

Figure 4.4.1: The three case-studies in the extrinsic motivation learning

The learning scheme is evaluated first in an illustrative setup with 10-DoF planar manipulator. First, the robot tries to reach some goals $\mathcal{G}_{train}$ driven by intrinsic motivation to gain sufficient knowledge about its internal model. After the initial exploration phase, if a human demonstration is observed and new novel goals are detected, the exploration is then guided utilizing the extrinsic signal (cf. Sec. 4.3.1). The robot is motivated to learn the detected novel goals as well as to discover novel outcomes and expand its knowledge about the workspace.

Three different case-studies are conducted to test the learning scheme on:

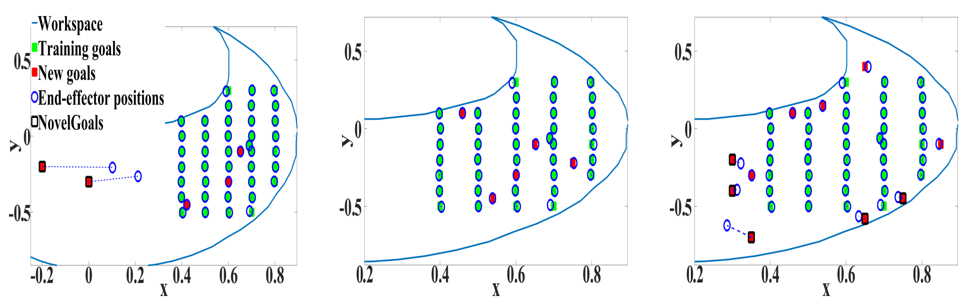
(A) case (1): $\mathcal{G}_{new}$ is scattered in the entire workspace (i.e, inside and outside the learned workspace) as illustrated in Fig. 4.4.1(a).

(B) case (2): $\mathcal{G}_{new}$ is scattered inside the learned workspace as illustrated in Fig. 4.4.1(b).

(C) case (3): $\mathcal{G}_{new}$ is scattered near the border of the learned workspace as illustrated in Fig. 4.4.1(c).

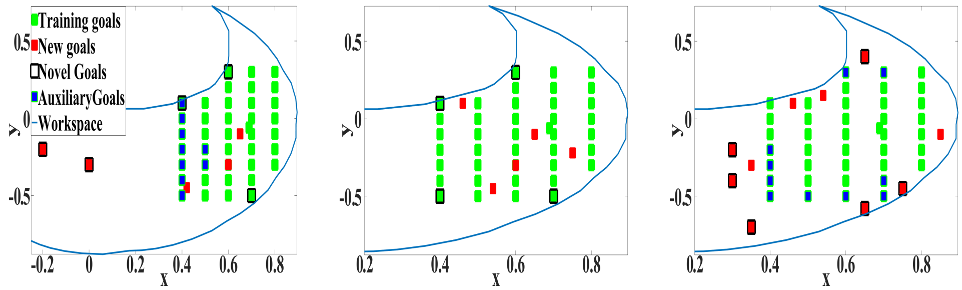
The blue dots in Fig. 4.4.1 represent the discovered workspace during the intrinsic motivation learning, the green dots represent $\mathcal{G}_{train}$, and the red ones represent $\mathcal{G}_{new}$ which is detected from simulated human demonstrations of goal locations.

The robot performance on $\mathcal{G}_{new}$ is first tested in each case, and $\mathcal{G}_{novel}$ is detected utilizing the novelty detection $\mathcal{N}$ (cf. Eq. (4.6)), where E_{thr} is determined automatically as the Max value, i.e., the upper inner fence UIF, of the performance error distribution obtained in the intrinsic motivation learning as explained in Sec. 4.2.2. The robot generalized and performed very well on the new goals which are located inside the learned workspace as illustrated in Fig. 4.4.2, where the blue circles represent the end-effector positions. None of these goals has been detected as a novel goal (cf. Fig. 4.4.3) and no additional learning is required. The robot also generalized well on some of the $\mathcal{G}_{new}$ which are located near the border of the learned workspace in case (3) (cf. Fig. 4.4.2(c)). The auxiliary goals $\mathcal{G}_a$ (cf. Sec. 4.3.1) are set only for the new novel goals. If some goals $g_i \in \mathcal{G}_{train}$ are not learned well, they are detected as novel goals but no auxiliary goal is set for them (cf. Fig. 4.4.3(b)). Note also that the robot performance on the detected novel goals is reasonable, even in case (3) (cf. Fig. 4.4.2(c)), due to LLM extrapolation which does not yield unpredictable behavior as discussed in Sec. 3.4.



(a) case (1) (b) case (2) (c) case (3)

Figure 4.4.2: The robot performance in the three case-studies

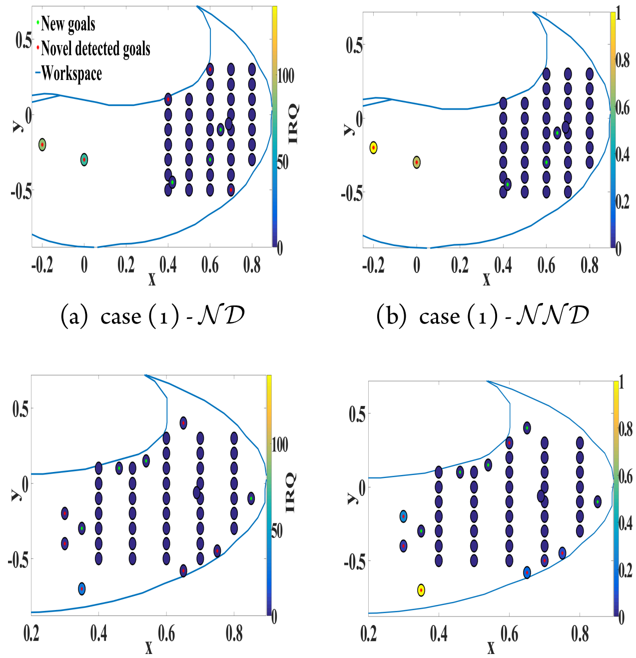


(a) case (1) (b) case (2) (c) case (3)

Figure 4.4.3: The goal sets in the three case-studies before extrinsic motivation learning

Fig. 4.4.4 shows the difference between the novelty degree $\mathcal{ND}$ and the normalized novelty degree $\mathcal{NND}$ (cf. Fig. 4.2.3). $\mathcal{NND}$ assigns (1) to the goal with the highest novelty degree in each of case (1) and case (3) (marked in yellow). The novelty degrees are $\mathcal{ND} = 118$ IQR for the corresponding goal in case (1), and $\mathcal{ND} = 40$ IQR for the corresponding goal in case (3). While $\mathcal{NND}$ assigned the same novelty measure to these goals in the two cases, $\mathcal{ND}$ assigned two different measures to them which reflects the actual robot's knowledge about the goals, meaning that the robot has more knowledge about the goal in case (3) than the one in case (1). Yet, $\mathcal{NND}$ gives comparable measures in each setup to drive the exploration.

The robot tries to reach the detected novel goals driven by the probabilistic extrinsic signal (cf. Sec. 4.3.1) until the robot gains enough knowledge about them, i.e., these goals are not detected as novel goals any longer ($E(g_i) \leqslant E_{thr}, \forall g_i \in \mathcal{G}$ (cf. Eq. (4.6))). Note that the robot learns not only the new goals but also the workspace scattered between $\mathcal{G}_{train}$ and $\mathcal{G}_{new}$. If the robot performance deteriorates (i.e., the performance error is larger than the performance error observed in the intrinsic motivation learning phase), this indicates that the robot potentially forgets about some previous experiences due to the continuous local model update. Consequently, the learning scheme (cf. Sec. 4.1) switches automatically to the intrinsic motivation learning mode, and the robot continues learning to en-



(a) case (1) - $\mathcal{ND}$ (b) case (1) - $\mathcal{NND}$

(c) case (3) - $\mathcal{ND}$ (d) case (3) - $\mathcal{NND}$

Figure 4.4.4: Novelty degree $\mathcal{ND}$ vs normalized novelty degree $\mathcal{NND}$

hance its performance on the full goal set.

The robot performance has been evaluated over 20 experiments in each case. The experimental results for 20 repetitions are illustrated in Table. 4.4.1. First, the robot explores driven by intrinsic motivation and tries to reach $\mathcal{G}_{train}$ within 500 epochs, each epoch consists of 100 time steps (samples). The robot performance is evaluated after each epoch.

After the intrinsic motivation learning phase, the robot gained some knowledge about its internal model with a validation RMSE of 3.9 mm and test RMSE of 1.9 mm. The robot demonstrated robust performance over 20 experiment illustrated with the RMSE std of 1.1 mm.

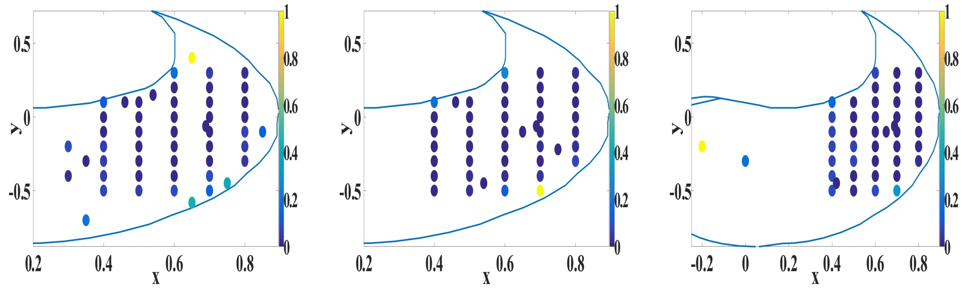
After the extrinsic motivation learning in each case, each of RMSE and RMSE std were improved significantly as illustrated in Table. 4.4.1. Case (1) required more learning epochs in order to discover a wider range of the workspace speci-

Table 4.4.1: The experimental results for the extrinsic-intrinsic motivation learning

Intrinsic motivation learning phase			
avg. RMSE validation [m]	$3.9 \cdot 10^{-3}$		
avg. RMSE std [m]	$1.1 \cdot 10^{-3}$		
avg. RMSE Test [m]	$1.9 \cdot 10^{-3}$		
Nr. Epochs	500		
Extrinsic motivation learning phase			
	case (1)	case (2)	case (3)
avg. RMSE validation [m]	$2.7 \cdot 10^{-3}$	$2.7 \cdot 10^{-3}$	$3.2 \cdot 10^{-3}$
avg. RMSE std [m]	$0.3 \cdot 10^{-3}$	$0.3 \cdot 10^{-3}$	$0.4 \cdot 10^{-3}$
avg. RMSE Test [m]	$1.9 \cdot 10^{-3}$	$1.9 \cdot 10^{-3}$	$1.9 \cdot 10^{-3}$
avg. Nr. Epochs	260	120	220
Additional intrinsic motivation learning phases	1\20	0	4\20

fied with the novel goals' location. Despite the reasonable performance accuracy in case (2), where no new novel goals were detected but rather some of the difficult-to-reach goals $g_i \in \mathcal{G}_{train}$, additional epochs were performed autonomously to enhance the performance accuracy on these goals because of the small E_{thr}. The smaller E_{thr}, the more samples are required, i.e. there is a trade-off between E_{thr} (i.e., the accuracy) and the efficiency. This was observed mainly in case (2) and case (3) where more learning epochs were performed to enhance the performance accuracy on the goals near the border.

Note that the aim of the extrinsic motivation learning is to expand the robot's knowledge about the workspace while maintaining the previously learned experiences. Therefore, the test RMSE does not necessarily improve further in this learning phase, as the robot focuses on learning novel outcomes and discover novel areas of the workspace and not on enhancing the previously learned workspace. In addition, the robot in the case (2) focuses on enhancing its performance on the border of the workspace. Still, the test RMSE is comparably good as in the intrinsic motivation learning phase, and the robot manages to generalize well on new goals in a wider range of the workspace.



(a) case (1) (b) case (2) (c) case (3)

Figure 4.4.5: The robot's interest in the extrinsic motivation learning

Fig. 4.4.5 illustrates the overall robot's interest in each goal during the extrinsic motivation learning phase, which is measured by how many times the robot tries to attain the corresponding goal in each case. The detected novel goals receive the highest interest of the robot indicated with yellow, green, and very light blue dots. The dark blue dots indicate that these goals are barely visited during this learning phase, they represent the goals in $\mathcal{T}$ which are chosen with very low probabilities ρ_0 (cf. Sec. 4.3.1). The azure dots (light blue dots in the training goal set) indicate that these goals have been visited more than the goals in $\mathcal{T}$ with higher probabilities $\rho_a > \rho_0$, as they represent the auxiliary goals as explained in Sec. 4.3.1.

4.5 Extrinsic-Intrinsic Motivation Learning with a Physical 7-DoF Baxter

The extrinsic-intrinsic motivation learning scheme with OEMR (cf. Sec. 3.5) has been implemented on the 7-DoF left arm of a physical Baxter robot (cf. Fig. 3.6.1). The parameter set is $\{\eta = 0.0725, \sigma = 0.0452, r = 0.0869, \lambda = 0.5, a = 4\}$, where η is the learning rate, σ is the exploratory noise (cf. Sec. 3.3), r is the radius parameter of LLM [32, 35, 44] (cf. Appendix .1), λ is the weighting factor of the interest signal (cf. Sec. 3.2), and a is the importance factor of the probabilistic intrinsic signal (cf. Sec. 4.3.2).

Intrinsic Motivation Learning Phase

First, the robot needs to gain some knowledge about its workspace driven by intrinsic motivation. To this aim, a similar experiment to the interest-driven exploration experiment with Baxter in Sec. 3.6 has been conducted. Note that a different three-dimensional virtual goal grid has been generated around the home position with 54 goals, and the robot's exploration is driven by the probabilistic intrinsic signal (cf. Sec. 4.3.2). Also, a different trajectory generator is used. Baxter is controlled by passing estimated joint values to a simple custom joint trajectory generator utilizing quintic polygons.

The goals are scattered in a cuboid shape, with a 10 *cm* vertical and horizontal distance between every two adjacent goals. These goals are used for the exploration in the real robot experiment, and they are visualized in MATLAB as illustrated in Fig. 4.5.2.

The robot starts exploring from its home position trying to reach the virtual goals. The goals are selected iteratively utilizing the probabilistic intrinsic signal. N intermediate targets are generated between each two selected goals, which varies based on the distance between these goals. After each epoch, which consists of 1000 time steps (samples), OEMR is performed and the robot performance is tested on the virtual goals. Each epoch took on average 75 *min* with a sampling rate of 3 sec per sample (i.e., per time), including training, performing OEMR, and evaluating the robot performance on the virtual goal set. Note that the virtual grid size is larger than the one generated in the previous experiment in Sec. 5.6.

The error converged very fast already after the first epoch as illustrated in Fig. 4.5.1. The robot performance accuracy was already good after 4 epochs (RMSE of 8.4 *mm*). Additional two epochs were performed to enhance further the performance error (RMSE of 7.6 *mm*) as well as to show the stability of the learner. Note that most of the goals were reached with an accuracy of less than 6 *mm*. Only five goals have a larger position deviation of more than 1 *cm* as they are located close to the limit of the robot's workspace.

The robot performance generalization was tested on 36 new goals randomly

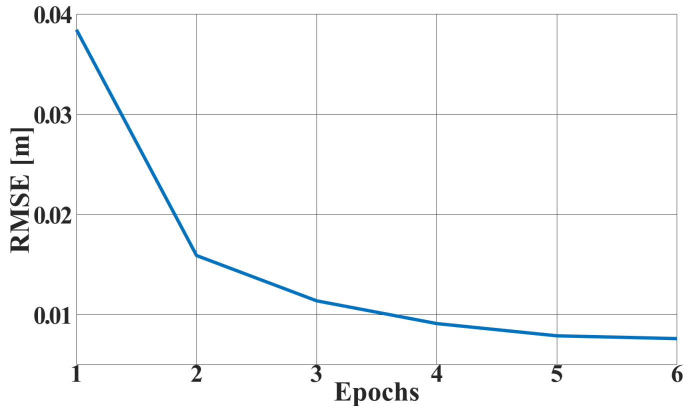


Figure 4.5.1: Baxter performance RMSE in the intrinsic motivation learning

scattered in the discovered workspace. The robot managed to reach them with an RMSE of 8.2 *mm*. Most of the goals were reached with accuracy less than 7.8 *mm*. Only three goals have a larger position deviation as they are located close to the limit of the robot's workspace.

Extrinsic Motivation Learning Phase

After the robot gained sufficient knowledge about its internal model and the workspace, three new goals are detected from observing human demonstrations' outcomes (i.e., the human reached three objects in front of the robot). One goal is located outside the virtual grid with a distance 28 *cm* to the learned workspace, one is located inside the grid, and one is near the border of the discovered workspace with a distance 15 *cm* to the learned workspace (cf. Fig. 4.5.2), in order to evaluate the learning scheme in the three case-studies (cf. Sec. 4.4).

First, the robot performance is evaluated on the new goals. The robot reached the goal inside the discovered workspace with an accuracy of 7.6 *mm*. The goals outside the learned workspace were reached with an accuracy of 4 *cm* and 6.65 *cm*, respectively, which increased the overall RMSE from 7.6 *mm* to 1.27 *cm* as illustrated in Fig. 4.5.4. Still, the robot performance on the new goals is reasonable and the robot did not demonstrate any unpredictable behavior as illustrated in

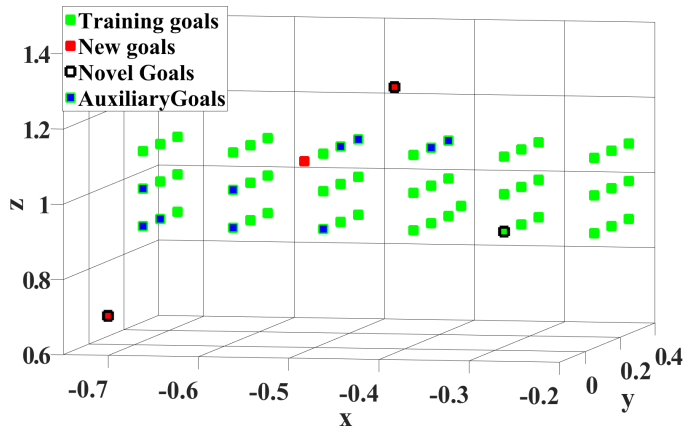


Figure 4.5.2: The goal sets after detecting new goals from observing human demonstrations' outcomes

Fig. 4.5.3. This indicates clearly the good extrapolation behavior of the LLM due to its incremental construction and the initialization of the added local models [32, 44] (cf. Appendix .1). The green dots in Fig. 4.5.3 represent the training goals which are generated in the intrinsic motivation learning, the red ones illustrate the new goals which are detected from observing human demonstrations' outcomes, and the blue circles represent the observed end-effector positions of Baxter.

Three novel goals are detected utilizing the novelty detection $\mathcal{N}$ (cf. Eq. 4.6):

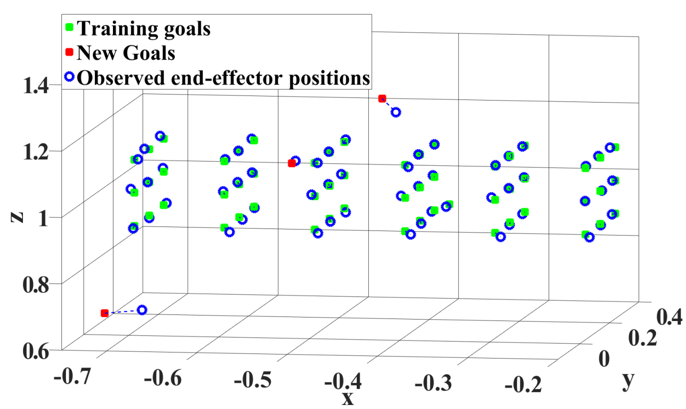


Figure 4.5.3: Baxter performance evaluation on the new goals

the goals outside the learned workspace and one of the difficult-to-reach goals from the training goal set. The auxiliary goals are generated only for the new novel goals as illustrated in Fig. 4.5.2.

Only two additional epochs were performed to learn the new goals with reasonable accuracy of 7.5 *mm* (RMSE) and discover the detected workspace between the newly induced goals and the training ones. The error converged very fast already after the first epoch. The error after the first epoch was reduced from RMSE of 1.27 *cm* to RMSE of 8.4 *mm* as illustrated in Fig. 4.5.4. The peak is due to the test on the three new goals before training.

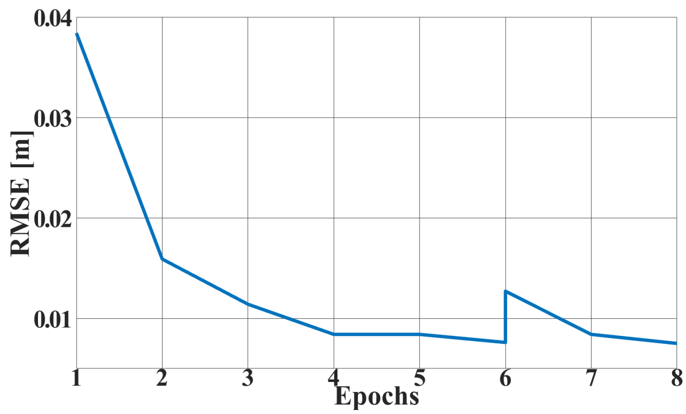


Figure 4.5.4: Baxter performance RMSE during the entire experiment. The first 6 epochs were performed during the intrinsic motivation learning, and the last two epochs were performed in the extrinsic motivation learning. The peak is due to the detected novel goals before the extrinsic motivation learning

Fig. 4.5.5 illustrates the discovered workspace during the intrinsic and the extrinsic motivation learning. The robot was motivated to discovered additional novel outcomes between the learned workspace in the intrinsic motivation learning and the novel goals. The light blue dots represent the discovered workspace in the extrinsic motivation learning and the dark blue ones represent the discovered workspace in the intrinsic motivation learning.

The robot performance is tested on new 64 goals scattered randomly in the en-

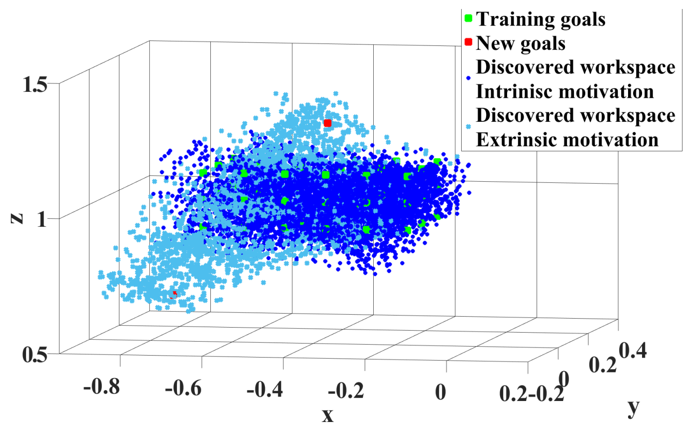


Figure 4.5.5: The discovered workspace during the intrinsic motivation learning (dark blue dots) and the extrinsic motivation learning (light blue dots)

tire discovered workspace. The robot managed to reach them with a performance accuracy of 7.8 *mm*. Most of the goals were reached with an accuracy of 6 *mm* and only a few with 8 *mm* around the border of the discovered workspace. The overall performance accuracy is very good compared to Baxter positioning accuracy of 5 *mm*.

The experiment demonstrated clearly the applicability of the proposed methods, their sample-efficiency, and the fast learning and adaptability. Only a few hours of direct online training were required to approximate the internal model with reasonable accuracy. Only two epochs were performed to learn and adapt fast to the newly detected goals and discover additional novel outcomes. The volume of the newly learned workspace is approximately similar to the one learned during the intrinsic motivation learning. The experiment also demonstrated that the probabilistic intrinsic signal leads to a comparable performance and accuracy similar to the interest measurement (cf. Sec. 3.6) in a real-world experiment. Moreover, the robot was able to decide on its own whether it needs to explore and learn further to achieve humans' goals, or it needs to exploit its prior-knowledge which is gained in the intrinsic motivation learning. The robot was able to benefit from observing human demonstration to expand its knowledge and improve its skills.

Note that the purpose of the real-world experiment is to evaluate each learning phase with the proposed methods. Therefore, two learning phases: intrinsic and extrinsic motivation learning are demonstrated. However, the learning scheme (cf. Sec. 4.1) can easily switch between intrinsic and extrinsic motivation autonomously. The initial first intrinsic motivation learning is necessary for the robot to gain some knowledge about its internal models. After that during the exploration, if a human demonstration is observed and a novel goal is detected, the learning is set automatically to the extrinsic motivation. Otherwise, the robot can continue exploring driven by intrinsic motivation.

4.6 Conclusion

This chapter designed a novel extrinsic-intrinsic motivation learning scheme which combines intrinsic motivation with learning from observation. There is hardly any research to integrate these two methods in the literature. Therefore, there is no direct computational model to compare it with the proposed framework so far.

The chapter adopted and adapted the observational learning concept as an extrinsic motivation learning. To this aim, three new methods are devised: novelty detection, novelty degree, and a probabilistic goal selection strategy. These methods provide together sample-efficient learning by selecting the most informative (novel) goals to learn from. Only two additional epochs were performed to adapt the model and learn the newly detected goals from observing human demonstrations' outcomes with reasonable accuracy. The robot also discovered a wider range of the workspace and learned novel outcomes not only the newly detected goals leveraging the probabilistic goal selection strategy and the auxiliary goals.

Despite the instantaneous update of the extrinsic and intrinsic learning signals, the robot demonstrated robust performance and consistent interest over 20 experiments as well as in direct online training on a physical robot. Utilizing the novelty methods, the robot was able to decide on its own which exploration strategy to follow and was able to manage the exploration-exploitation trade-off. All the goals were reached with reasonable accuracy within only a few training epochs.

How can Goal Babbling learn inverse models, e.g., inverse kinematics, with multiple solutions for redundant robots online and in a stable fashion?

5 Associative Goal Babbling for Redundant Robots

The underlying exploration method in the proposed learning schemes in the previous chapters (chapter. 3, chapter. 4) relies on Goal Babbling in order to leverage the advantages of Goal Babbling, e.g., direct inverse model learning, adaptability, learning while behaving, scalability to high DoF robots, and online learning from scratch (cf. Sec. 2.5, Sec. 3.3). For redundant robots, Goal Babbling by design [32] learns inverse models, e.g., inverse kinematics, with only one solution which restricts the flexibility of the robots. Associative dynamic networks [40, 41] have been proposed to tackle this challenge and provide a suitable representation for multiple solutions. However, the exploration with Goal Babbling and the solution consolidation in the network have been done separately. Moreover, the associative network in [40, 41] works only offline, which is incompatible with lifelong learning. The network has a fixed size without any online adaptation possibility. The full data set needs to be stored to train the network offline [40].

This chapter devises an Online Associative Radial Basis Function network (OARBF) to enable learning inverse models, e.g., inverse kinematics, with multiple solutions online with Goal Babbling for redundant robots. OARBF is constructed incrementally, from scratch, and updated continuously in order to gain more flexibility and adaptability through lifelong learning for developmental robots. In real applications, learning while behaving produces very noisy data, which makes the dynamic network potentially unstable. To mitigate this effect and tackle the stability problems, this chapter establishes a parameter-sharing technique which assures stability by combining and synchronizing incremental regression with associative dynamics to leverage their advantages: stability, accuracy, and multi-model representations. This technique also increases sample-efficiency by drastically reducing the number of required training samples for OARBF and the dimensionality of the parameter space. It also speeds up the learning process by synchronizing two learners' updates.

OARBF is first compared to the offline ARBF in an illustrative 10-DoF planar manipulator in order to highlight its advantages. It is then integrated into a novel hierarchical interest-driven associative learning scheme and evaluated on a physical 7-DoF Baxter robot arm. The results highlight clearly the high stability and rapid adaptability of the network even in the presence of noisy data with direct online training on real robots. One of the learned solutions is selected automatically based on the previous state of the robot.

This work has been already published:

- R. Rayyes, H. Donat, and J. Steil, "Hierarchical interest-driven goal babbling for efficient bootstrapping of sensorimotor skills," in 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020, pp. 1336–1342. [31]

- R. Rayyes and J. Steil, "Online associative multi-stage goal babbling toward versatile learning of sensorimotor skills," in 2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2019, pp. 327–334. [132]

Moreover, it has been also presented at the following workshops:

- R. Rayyes and J. Steil, "Hierarchal interest-driven associative goal babbling", The Annual Conference on Neural Information Processing Systems (NeurIPS), WiML workshop, Vancouver, Canada, 2019.
- R. Rayyes and J. Steil, "Interest-based exploration with associative goal babbling toward versatile cognitive robots", Robotics Science and Systems (RSS): WiR workshop, Freiburg, Germany, 2019.
- R. Rayyes and J. Steil, "Interest-based exploration with associative goal babbling", The Fourth International Workshop on Intrinsically Motivated Open-ended Learning (IMOL), Frankfurt, Germany, 2019.

5.1 Online Associative Radial Basis Function Network

OARBF is constructed incrementally, from scratch, and updated on the fly. Its complexity (i.e, the hidden layer size) is adapted to the learned problem autonomously and continuously.

The basic structure of OARBF is similar to ARBF [40]. However, it is fundamentally constructed differently. As illustrated in Fig. 5.1.1, OARBF consists of three layers: an input layer, an output layer and one hidden layer of radial basis functions h. The input and the output layers are usually identical with the same number of neurons, each consists of $n+m$ neurons which corresponds to $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$ respectively, where n and m are the input data dimensions. For example, the input (inp) and output (out) data for learning kinematics are vectors that concatenate the end-effector positions $a = x \in \mathcal{X} \subset \mathbb{R}^n$ and configurations $b = q \in \mathcal{Q} \subset \mathbb{R}^m$ $(inp = out = [x, q]^T \in \mathbb{R}^{n+m})$. n is the dimension of the

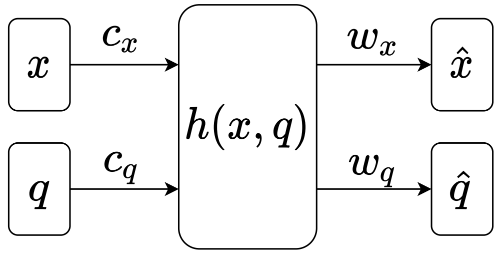


Figure 5.1.1: The general structure of the associative radial basis function network

end-effector position x, and m is the dimension of the configuration space, i.e., # DoF.

The input data should be normalized in order to have equivalent contributions to the network activity and balance the distribution of the neurons. c_x and c_q are the centers of the radial basis functions corresponding to the inputs x and q respectively. w_x and w_q are the weights of the output layer corresponding to the outputs x and q respectively.

The neurons in the hidden layer are added incrementally during learning based on the online data stream. The network is initialized with the first neuron centered around the first received sample $(x_1 = x^{home}, q_1 = q^{home})$, i.e., $\{c_{x_1} = x_1, c_{q_1} = q_1\}$. The initial output weights are initialized with the inputs weights $\{w_{x_1} = c_{x_1}, w_{q_1} = c_{q_1}\}$ in order to shift the output of the network to the current first sample. When the network receives a new sample (x_{new}, q_{new}) which has a distance larger than a radius r to all existing neuron centers, a new neuron will be added and centered around the newly received sample $\{c_{x_{(i+1)}} = x_{new}, c_{q_{(i+1)}} = q_{new}\}$. The output weights of this neuron are initialized with the output weights of its closest neuron neighbor to avoid drastic changes in the learned function. The neurons are organized and updated continuously based on Instantaneous Topological Map (ITM) [133].

The basis function activity for each neurons i is given in Eq. (5.1). The association setup solves the ambiguity of the inverse kinematics by utilizing mixed representations in the hidden layer. Hence, the data pairs (x_k,q_k), (x_l,q_l) have different representations in the hidden layer $h(x_k, q_k) \neq h(x_l, q_l)$, where $x_k = x_l$, $q_k \neq q_l$, $k \neq l$ [40].

$$\left.\begin{aligned} h_i(x, q) &= \text{softmax}(a_i(x,\ q)) = \frac{a_i(x,\ q)}{\sum_{j=1}^{H} a_j(x,\ q)} \\ a_i(x,\ q) &= \exp(-\beta_x d(x, c_{x_i})^2 - \beta_q d(q, c_{q_i})^2) \end{aligned}\right\} \tag{5.1}$$

d is the Euclidean distance between the newly received sample and center of the neuron i, H is the number of the hidden neurons which are added incrementally, softmax scales the sum of activation functions a to unity, and β_x and β_q are used to control the spread as well as the overlap of the radial basis functions, and is given in Eq. (5.2):

$$\beta_q = \frac{n}{m}\beta_x \tag{5.2}$$

n and m are the input data dimensions.

The output of OARBF is estimated based on the hidden layer's activation as follows:

$$\hat{out}(x, q) = w^{out} h(x, q) \tag{5.3}$$

The output weights w^{out} are updated at each step by performing a gradient descent in order to minimize the weighted error E_t (cf. Eq. (5.4)):

$$E_t = w_t^{gb} \|inp_t - \hat{out}_t\|^2 \tag{5.4}$$

$$w_{t+1}^{out} = w_t^{out} - \eta \cdot \frac{\partial E_t}{\partial w_t^{out}} \tag{5.5}$$

$\hat{out}$ is the estimated output of OARBF (cf. Eq. (5.3)), w^{gb} is the weight of the data sample (cf. Sec. 3.3, Eq. (3.4)), η is a learning rate, t is a time step, and $w^{out} = [w_x, w_q]^T$.

Note that the equations Eq. (5.1) and Eq. (5.3) are similar to the equations in [40]. However, the output weights w^{out} are initialized and updated differently. In addition, the hidden layer size H is not fixed but rather continuously updated as

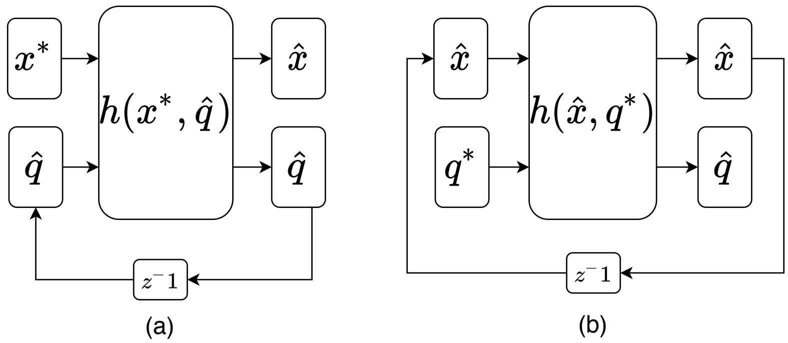
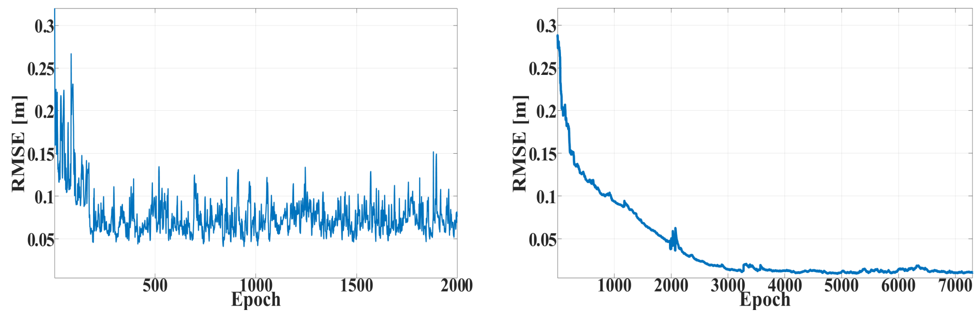


Figure 5.1.2: Output feedback driven loop to query OARBF, (a) for inverse kinematics: the network is driven by the desired x_t^*, and the last $\hat{q}_{t-1}$ is fed back into the network, (b) for forward kinematics: the network is driven by the desired q_t^*, and the last $\hat{x}_{t-1}$ is fed back into the network. The output feedback loop is iterated until convergence

OARBF is constructed incrementally from scratch. Moreover, only β_x needs to be tuned in contrast to [40].

Inverse kinematics (IK) as well as forward kinematics (FK) are learned simultaneously utilizing OARBF. An output feedback-driven loop is established (cf. Fig. 5.1.2) to query the learned model. The network converges to one of the learned solutions based on the previous state of the network. To query IK of the trained OARBF for example, the network is driven by the desired x_t^* and the last $\hat{q}_{t-1}$ is fed back into the network. The output feedback loop is iterated until convergence (cf. Fig. 5.1.2(a)), i.e., $E_t^x \leqslant E_{th}$, where $E_t^x = w_t^{gb} \|x_t^* - \hat{x}_t\|^2$ and E_{th} is an error threshold. To query FK of the trained OARBF, the network is driven by the desired q_t^* and the last $\hat{x}_{t-1}$ is fed back to the network. The output feedback loop is iterated until convergence (cf. Fig. 5.1.2(b)), i.e., $E_t^q \leqslant E_{th}$, where $E_t^q = w_t^{gb} \|q_t^* - \hat{q}_t\|^2$. Hence, the output feedback is added in order to select one of the learned solutions based on the previous state of the network to avoid inconsistencies. For example, OARBF selects an elbow-up or an elbow-down configuration to reach a desired position for a 2 DoF manipulator based on the given previous robot state.



(a) Unrobust and unstable performance

(b) Inefficiecnet learning after exhuastive parameter-tuning

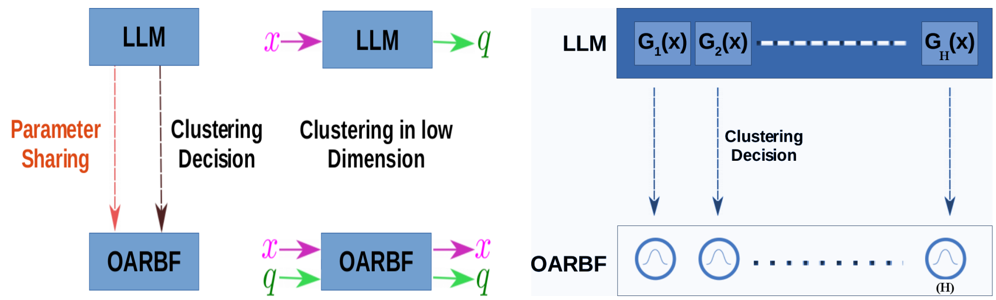
Figure 5.2.1: The challenges of OARBF in practical experiments

5.2 Parameter-Sharing Technique

OARBF has been tested with a 10-DoF planar manipulator (cf. Fig. 5.3.1). The main challenges of this network are stability and sample-efficiency.

OARBF demonstrated highly unstable performance as illustrated in Fig. 5.2.1(a). In online learning, the exploration is informed by the learning and vice versa [134], i.e., the network is trained and exploited continuously at each learning step. However, the network, in the beginning, yields an estimated output with very high error since the model complexity is still too simple to approximate the learned problem. Accordingly, the feedback loop in the exploitation phase can lead to error amplification when the network is output feedback-driven [42]. The informed exploration step thus is based on a highly amplified error. In addition, the high error of OARBF, as well as the high dynamical system resulting from the incremental constructing of the network, from continuous establishing, and from removing the feedback loop for exploiting and training OARBF, leads to unstable performance as well as the inhomogeneous distribution of the neurons.

Even when the parameters are well-tuned (e.g., utilizing pattern search optimization [135]) and a regularization is added, OARBF has demonstrated inefficient performance and needed a long time to converge as shown in Fig. 5.2.1(b).

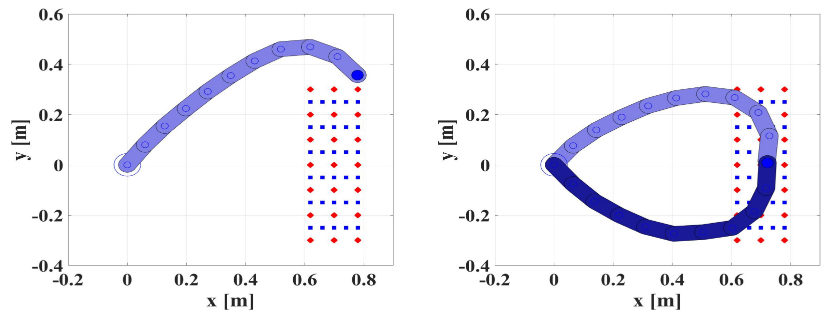


(a) Parameter-sharing technique (b) Sharing the clustering decision

Figure 5.2.2: Parameter-sharing technique: online clustering for LLM is done in the low dimensional Cartesian space. When a basis function of LLM $G(x)$ is added, the corresponding neuron of OARBF is added simultaneously, $H = \#$ LLM basis functions $= \#$ OARBF hidden neurons. One parameter set $\theta = \{\eta, r\}$ is shared with both learners

This chapter establishes a parameter-sharing technique to increase sample-efficiency, speed up the learning process, and stabilize the full learning system. The parameter-sharing technique combines incremental regression with associative dynamics by connecting two learners (LLM and OARBF) to leverage their advantages. LLM has demonstrated high stability and high accuracy in real robot experiments (cf. Sec. 3.6, Sec. 4.5, [39]) as well as for approximating complex models (e.g, [35]), while OARBF can model and represent multiple solutions by using multi-stable dynamic attractors.

Both learners are constructed online, incrementally, and from scratch. Hence, online clustering of the input data is needed to add the prototypes of LLM and the neurons of OARBF incrementally (cf. Sec. 5.1, Appendix .1). The input of LLM for learning IK is $x \in \mathbb{R}^n$ and the output is $q \in \mathbb{R}^m$. While OARBF input and output data is $[x, q]^T \in \mathbb{R}^{n+m}$. Accordingly, the online clustering for LLM is performed in the low dimensional input space (n D), while it is performed for OARBF in high dimensional input space ($(n + m)$ D). Therefore, in the parameter-sharing technique, the online clustering is done only for LLM in the low dimensional Cartesian space. When the prototypes of LLM are added, the neurons of OARBF are added simultaneously (cf. Fig. 5.2.2). This accelerates the clustering and stabilizes the full



(a) Training and test goal sets

(b) Reaching the same position with two different configurations

Figure 5.3.1: 10-DoF planar manipulator

learning system as it yields a more homogeneous distribution of the prototypes and the neurons.

Moreover, only one parameter set $\theta = \{\eta, r\}$ is shared with both learners, where η is the learning rate and r is the radius which determines the vicinity of each basis function (cf. Sec. 5.1, Appendix .1). $\beta_q = \frac{n}{m}\beta_x$ (cf. Eq. (5.1)). Consequently, only 3 parameters $\{\eta\ ,\ r\ ,\ \beta_x\}$ need to be tuned for both learners instead of 6 : $\{\eta_{LLM}\ ,\ \eta_{OARBF}\ ,\ r_{LLM}\ ,\ r_{OARBF}\ ,\ \beta_x,\ ,\ \beta_q\}$, and only one clustering is performed. Accordingly, utilizing parameter-sharing technique drastically reduces the dimensionality of the parameter space and synchronizes the learners'update.

5.3 OARBF with Parameter-Sharing Technique Evaluation

OARBF has been implemented with the original Goal Babbling [32] and parameter-sharing technique for learning the kinematics of a 10-DoF planar manipulator (cf. Fig. 5.3.1, each link length is 10 cm). This is the same experimental setup as the one done for testing OARBF in Sec. 5.2.

21 predefined goals regularly distributed on a grid are used for training and 30 goals are used for testing. Fig. 5.3.1(a) illustrates the training (red squares) and

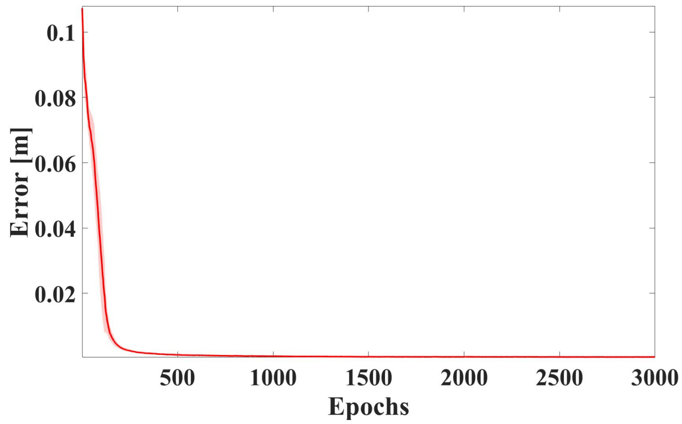
test (blue squares) goals. Each goal can be reached with a curvature-up configuration as well as with a curvature-down configuration (cf. Fig. 5.3.1(b)). The data is normalized to $[-1,\ 1]$ given the joint limits and the task dimension.

The robot starts exploring from its home position x^{home} with a curvature-up home posture ($q^{home} = [-0.3\pi, -2\pi/27, -2\pi/27, ..., -2\pi/27]^T\ rad$) trying to reach the training goals. After 3000 epochs, each consists of 100 time steps (samples), the robot switches its home posture to a curvature-down configuration ($q^{home}_{new} = -q^{home}$), and continues exploring trying to attain the same training goals. The learners are updated continuously and simultaneously at each time step, 66 neurons were added incrementally to OARBF during learning with ($r = 0.08, \eta = 0.01, \beta_x = 5, \beta_q = 1$).

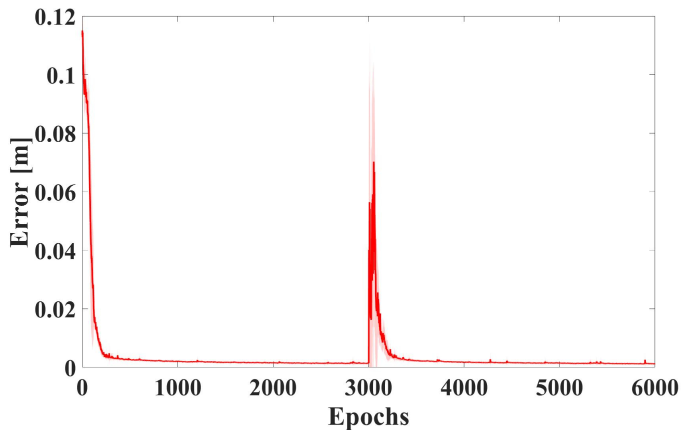
The robot performance is evaluated on the training and test goal sets with two different initial starting configurations. The robot manages to reach all goals without any inconsistencies, i.e., all goals are attained with curvature-up or curvature-down configurations based on the initial robot state.

The experiment has been repeated 10 times. The training errors for both learners converge very fast already after 200 epochs as illustrated in Fig. 5.3.2. The longer training time is to show the stability of the learners. The RMSE of OARBF increases when the robot switches its home posture, which is indicated with the peak in Fig. 5.3.2(b). The shaded area in the figure represents the std RMSE over 10 experiment runs. The small std indicates the robust performance of OARBF over the experiments.

Only 200 training epochs were required for OARBF, in contrast to the previous experiment with OARBF in Sec. 5.2 where the performance error of OARBF took 3000 epochs to converge (cf. Fig. 5.2.1(b)). Accordingly, utilizing the parameter-sharing technique reduced drastically the number of training samples 15 times for OARBF. In addition, Fig. 5.3.2(b) illustrates the high stability of OARBF due to the parameter-sharing technique in contrast to the previous experimenter (cf. Fig. 5.2.1(a)).



(a) *LLM*



(b) *OARBF*

Figure 5.3.2: Performance error - std RMSE of OARBF and Goal Babbling for a 10-DoF planar manipulator. The peak in (b) is due to the configuration switch

Table 5.3.1: OARBF with Goal Babbling results for the 10-DoF planar manipulator

		FK	IK		FK	IK
		$[m]$	$[m]$		$[m]$	$[m]$
Curve-up	LLM Trainig	N.A.	$1.7 \cdot 10^{-3}$	Curve-down	N.A.	$1.8 \cdot 10^{-3}$
	LLM Test	N.A.	$1.7 \cdot 10^{-3}$		N.A.	$1.7 \cdot 10^{-3}$
	OARBF Training	$1.3 \cdot 10^{-2}$	$6.5 \cdot 10^{-3}$		$1.3 \cdot 10^{-2}$	$5.7 \cdot 10^{-3}$
	OARBF Test	$2.4 \cdot 10^{-2}$	$7 \cdot 10^{-3}$		$1.5 \cdot 10^{-2}$	$6.1 \cdot 10^{-3}$

The forward and inverse kinematics are learned simultaneously. The average training and test RMSE are illustrated in Table 5.3.1; N.A. means not applicable. As illustrated in the table, LLM yields better accuracy than OARBF. This is due to several reasons. First, exploiting OARBF requires establishing an iterative feedback loop which might lead to error amplification [40]. In contrast, LLM is exploited directly, and its estimated output has therefore better accuracy without any error amplification. Second, according to the parameter-sharing technique, LLM is the main learner which is responsible for organizing the online clustering, and for the exploration. Hence, the shared parameters between the learners have been tuned for LLM and not of OARBF. In addition, the better accuracy of LLM is also due to LLM structure, i.e., approximating the learned function with locally weighted linear functions [44] (cf. Appendix. .1), as well as due to its lower dimensionality input and output (cf. Sec. 5.2). However, OARBF is able to learn multiple solutions and has demonstrated robust performance as well. Hence, the proposed system leverages their combination.

The accuracy of the forward model is slightly worse than the accuracy of the inverse kinematics. This is due to the spread of the basis function (β_x, β_q) [40]. These parameters are obtained utilizing pattern search optimization [135] to enhance the inverse kinematics trading the accuracy of the forward kinematics. This trade-off is favorable because acquiring reaching skills mainly relies on learning inverse kinematics in this thesis.

5.4 Comparison: Offline ARBF vs OARBF

In order to compare the performance of OARBF with offline ARBF, OARBF with Goal Babbling has been implemented in a similar experimental setup as in [40] with a 10-DoF planar manipulator (cf. Fig. 5.3.1). The robot needs to explore half of the workspace. The comparison results are:

1. The complexity of OARBF (i.e., the number of neurons) has been reduced and tailored to the problem in contrast to ARBF. 215 neurons are added

to OARBF incrementally to achieve approximately the same accuracy of 5.6 *mm* as in [40], where the network size of ARBF is pre-fixed with 300 neurons.

2. The exploration in [40] has been done first. The solutions are then consolidated in ARBF offline. In contrast, OARBF is trained and constructed online during exploration. Consequently, the consolidation of the solutions and the exploration have been done simultaneously.

3. ARBF in [40] is trained offline and the entire data set needs to be stored. In contrast, OARBF is trained online and updated on the fly. Therefore, no data set needs to be stored.

4. Only one online clustering is performed for both LLM and OARBF using the parameter-sharing technique (cf. Sec. 5.2). In contrast, an additional clustering phase is required in [40] to obtain c_x and c_q for ARBF. Therefore, the parameter-sharing technique speeds up the learning process.

5. Only 4 parameters need to be tuned in OARBF and Goal Babbling $(\sigma\ ,\ r\ ,\ ,\eta, \beta_x)$ because of the parameter-sharing technique (cf. Sec. 5.2). In contrast to [40], where 8 parameters needs to be tuned $(\sigma\ ,\ r_{LLM}\ ,\ r_{ARBF}\ ,\ \eta_{LLM}\ ,\ \eta_{ARBF}\ ,\ \beta_x\ ,\ \beta_q,$ #neurons$)$. Accordingly, the parameter-sharing technique drastically reduces the dimensionality of the parameter space to the half.

In order to evaluate OARBF in a real-world experiment, OARBF is first integrated with the interest-driven exploration scheme as explained in the next section.

5.5 Hierarchical Interest-Driven Associative Goal Babbling Scheme

The hierarchical interest-driven associative Goal Babbling scheme combines the interest-driven Goal Babbling with OARBF. The learning scheme enables the

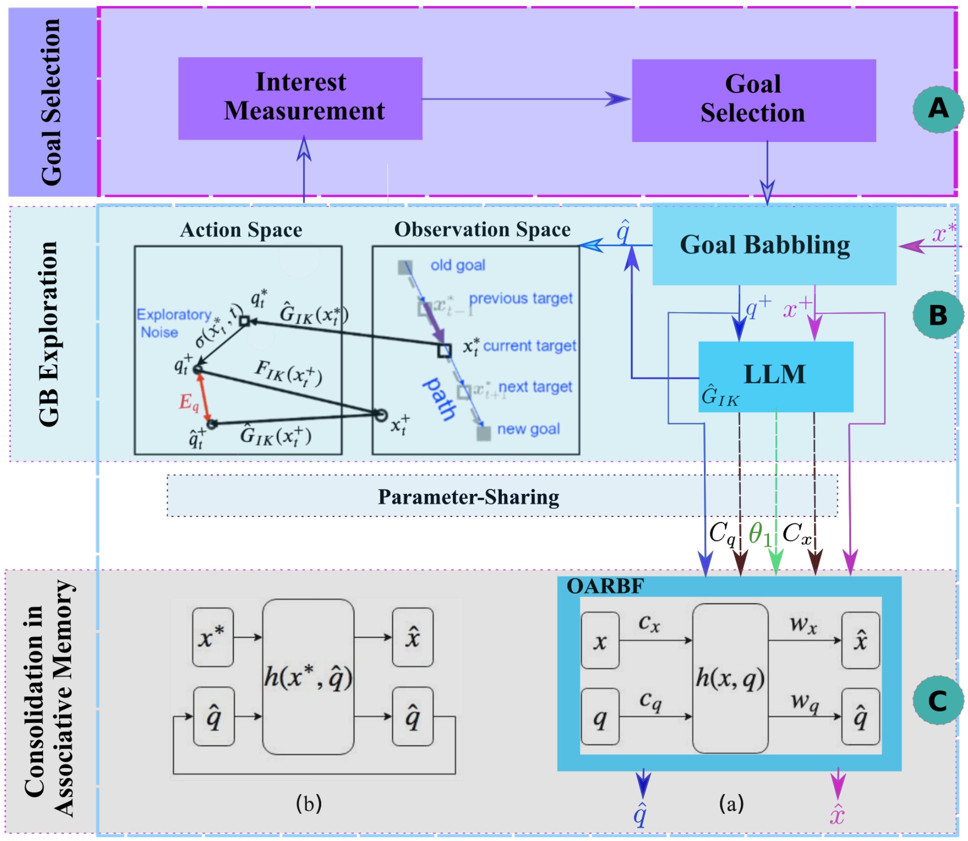


Figure 5.5.1: Hierarchical interest-driven associative Goal Babbling scheme. It comprises a high level of goal selection strategy and two low levels of exploration and solution consolidation. (A) Goal selection strategy, (B) Goal-directed exploration mechanism, (C): Associative dynamic memory (a) OARBF training, (b) establishing a feedback loop for OARBF exploitation

robot to explore driven by intrinsic motivation and to learn its models online in a learning while behaving fashion. The learning scheme also allows the robot to learn multiple solutions by using OARBF in order to accomplish required tasks flexibly.

Fig. 5.5.1 illustrates the general architecture of the learning scheme which consists of three levels:

(A) Goal selection: The goals are selected utilizing the interest measurement (cf. Sec. 3.2).

(B) Goal-directed exploration mechanism: The exploration relies on the interest-driven Goal Babbling (cf. Sec. 3.3) to permit learning while behaving fashion.

(C) Associative memory: OARBF is implemented to permit online learning multiple solutions of inverse kinematics with Goal Babbling.

The learning scheme is fully online and updated continuously on the fly. It does not require storing full data sets or any intermediate offline training.

5.6 Hierarchical Interest-Driven Associative Goal Babbling with a Physical 7-DoF Baxter

Hierarchical interest-driven associative Goal Babbling with OEMR (cf. Sec. 3.5) has been implemented on a physical 7-DoF left arm of the Baxter robot, (cf. Fig. 3.6.1). The inaccurate positioning accuracy of 5 mm of Baxter as well as the learning while behaving imposes additional challenges for the learners to cope up with the noisy data and for stabilizing the incremental OARBF.

The parameter set is $\{\eta = 0.0725, \sigma = 0.0452, r = 0.0869, \lambda = 0.5, \beta_x = 5\}$, where η is the learning rate, σ is the exploratory noise (cf. Sec. 3.3), r is the radius parameter of LLM [32, 35, 44] (see Appendix .1), λ is the weighting factor of the interest measurement (cf. Sec. 3.2), and β_x is used to control the spread as well as the overlap of the basis radial functions (cf. Sec. 5.1).

The task is here, the robot should learn to reach some desired positions in the workspace with two different configurations, without any prior-knowledge of the model, online, from scratch, in a learning while behaving fashion, and driven by its interest.

The exploration has been done in two phases with two different home postures heuristically chosen:

- $q_1^{home} = [-0.17, -0.25, -0.12, 0.93, -0.71, 1.72, 0.61]^T \, rad$

- $q_2^{home} = [-1.20, -0.615, 0.38, 1.34, 0.29, 1.28, -0.329]^T \; rad$

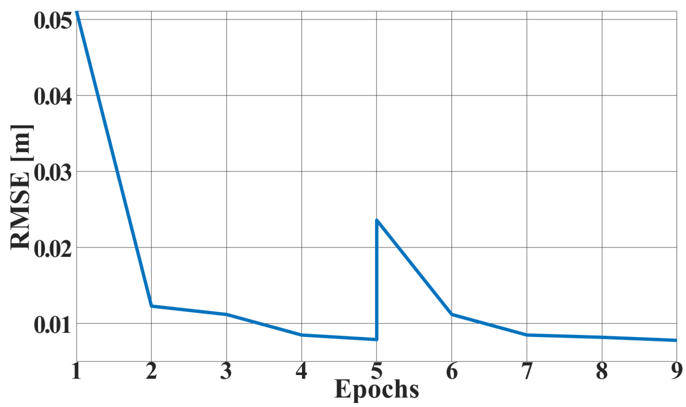


Figure 5.6.1: OARBF Performance RMSE. The peak is due to the configuration switch

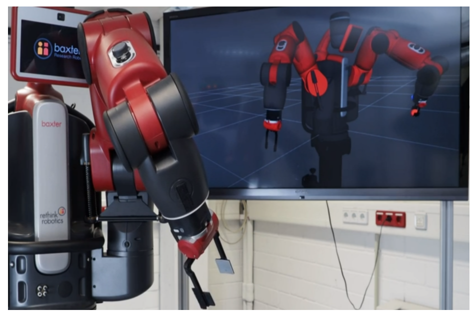
As described in Sec. 3.6, the experimenter shows the robot a desired workspace to be explored. A three-dimensional virtual goal grid of 30 goals is created inside the defined volume. The goals are scattered in a cuboid shape, with a vertical and horizontal distance of 10 cm between every two adjacent goals. Note that these virtual goals are used for the exploration with the physical robot (cf. Fig. 3.6.3). A supplementary video illustrating the experiment is available at `https://youtu.be/W6tB-7fos4A` [128].

The robot starts exploring from its home position x^{home} corresponding to q_1^{home} trying to reach these virtual goals. The goals are selected iteratively utilizing the interest measurement (cf. Sec. 3.2). Each goal trial consists of N intermediate targets. The number of intermediate targets varies depending on the distance between the selected goals. The robot performance is evaluated on the virtual goal set after each epoch (1000 samples, i.e., 1000 time steps) and OEMR is performed. After 5 epochs, the robot switches its home posture to q_2^{home} and continues exploring trying to reach the same virtual goals for another 5 epochs. Each epoch took approximately 60 min to complete.

(a) First home position

(b) Reaching with a first configuration

(c) Second home position



(d) Reaching with a second configuration

Figure 5.6.2: Baxter reaches a virtual goal with two different configurations based on the previous robot configurations

The error converged very fast already after the first epoch as shown in Fig. 5.6.1. The peak is due to the configuration switch. The robot demonstrated already a good performance after two epochs in each learning phase. However, additional three epochs were performed to illustrate the stability of OARBF. The training RMSE is 6.7 *mm*.

After training, the robot performance is evaluated on 27 new goals with two different initial starting configurations. The robot managed to reach all goals without any inconsistencies based on its initial configuration, i.e., without switching solutions or averaging between them as illustrated in Fig. 5.6.2 and in the video.

22 goals were reached with a test RMSE of 6.4 *mm*, and only 5 goals with an avg. RMSE of 8.6 *mm* as they were difficult to reach because of self-collision avoidance by the robot system. The achieved accuracy is acceptable compared to the low

accuracy of Baxter. The execution time for the iterative loop to convergence to one of the learned solutions was negligible as illustrated in the video.

The results demonstrated clearly the high stability of the OARBF even with direct online training on a real robot. Only 5 parameters needed to be set despite using two learners due to the parameter-sharing technique (cf. Sec. 5.2). The robot was able to learn the required internal model with reasonable accuracy and sample-efficiency demonstrated with the few required training epochs.

5.7 Conclusion

This chapter devised an incremental online associative radial basis function (OARBF) network. The network is constructed totally from scratch and incrementally. It is able to learn multiple solutions of inverse kinematics with Goal Babbling *online* using multi-stable attractor to replicate the versatile coordination observed in humans. The network demonstrated a robust performance without any inconsistencies by always converging to one of the learned solutions based on the initial robot state without switching or averaging solutions.

The chapter also established a parameter-sharing technique which stabilizes the full learning system by combining and synchronizing two learners. The learning system demonstrated high stability in a real-world application despite the presence of noisy data resulting from learning while behaving and the low accuracy of the robot, and despite the high dynamics resulting from the incremental OARBF construction and establishing its feedback loops. The learners are synchronized through performing only one clustering to add the basis functions of the learners and updating them simultaneously. Parameter-sharing technique reduces the efforts required for tuning the parameters by drastically reducing the dimensionality of the parameter space. It increases the sample-efficiency significantly by reducing the number of training samples required for OARBF to learn the model with reasonable accuracy at least 15 times.

The advantages of OARBF is highlighted compared to the offline ARBF: (i) The exploration and the consolidation of multiple solutions are done simultaneously;

(ii) the complexity of the network is tailored to the learned problem with continuous adaptability; (iii) the network is updated continuously on the fly; (iv) no training data is needed to be stored; (v) the dimensionality of the parameter space is halved due to the parameter-sharing technique.

The methods are integrated into a novel hierarchical associative learning system, which permits exploration in a learning while behaving fashion driven by the robot's interest, and learns several solutions to accomplish the required task *flexibly*. All the proposed methods are demonstrated with direct online training on a physical 7-DoF Baxter robot arm. They demonstrated high stability and high efficiency to learn the inverse kinematics model with two different solutions within a few hours of direct training.

The end is just a new beginning

6 Conclusion

6.1 Summary

This thesis focuses on devising efficient and stable online learning schemes for developmental robots with direct online training on real robots. Despite the plethora of interesting previous work in this field, there are three open research problems to be considered:

First, lifelong learning requires intrinsic motivation methods to guide the self-exploration of developmental robots, and requires online learning as an essential capability to provide adaptation. However, the high sample-complexity of online and intrinsic motivation methods renders direct online training on physical robots very challenging. Consequently, the majority of related works have been demonstrated only in simulation. It is therefore an open research question on how to devise efficient online intrinsic motivation methods for real-world applications.

Second, developmental robots often share the environment with humans, and learning from them accelerates the autonomous development of these robots. Still, there is hardly any research to integrate intrinsic motivation with observational learning. Combining these methods in one learning scheme imposes additional challenges for lifelong learning. Such a learning scheme must provide a suitable *exploration strategy* to autonomously choose between self-exploration and learning from observation. Besides, the robot should manage the *exploration-exploitation trade-off*. How to integrate intrinsic motivation with learning from observation and enable the robot to decide autonomously how and when to explore is, therefore, another open research question.

Third, the previously proposed Goal Babbling method provides desirable advantages for lifelong learning. It is thus promising to be considered as an efficient underlying exploratory method for devising such learning schemes. However, learning inverse models with multiple solutions *online* with this framework was missing and highly desirable for high DoF redundant robots. One way to tackle this challenge is to combine Goal Babbling with associative dynamic networks. However, online incremental dynamic networks impose *stability* challenges and bring up the question of how to learn several solutions for redundant robots with Goal Babbling *online* with *stability* in the presence of noisy data in real applications.

These three research questions motivated this thesis to tackle the challenges behind devising efficient and stable online learning schemes for developmental robots for real-world applications. To this aim, the following contributions have been obtained:

1. A novel intrinsic motivation method named "interest measurement" is established in chapter 3. New intrinsic motivation signals: knowledge-based signal called "relative error" and competence-based signal called "forgetting factor" are devised and unified in the interest measurement method so that the robot is able to: (i) learn from the most informative goals and generalize on simpler ones in order to increase sample-efficiency by utilizing the relative error; (ii) autonomously focus again on forgotten, previously learned

goals in order to assure the lifelong learning concept by utilizing the forgetting factor. In addition, a novel online mental replay method named ”online episodic mental replay (OEMR)” is devised to intensify the robot’s experiences and facilitate deploying online data-driven learning methods on real robots. The proposed online mental replay method does not require storying full data sets or augmenting the goal space, in contrast to other state-of-the-art replay methods.

2. A novel extrinsic-intrinsic motivation learning is established in chapter 4 which integrates observational learning with intrinsic motivation. It enables the robot to explore driven by intrinsic motivation, as well as to benefit from observing human demonstrations’ outcomes in order to expand its knowledge about the workspace and accelerate learning. To this aim, three new methods are devised: Novelty detection, novelty degree, and a probabilistic goal selection strategy. The novelty methods enable the robot to decide on its own which exploration strategy to follow: intrinsic or extrinsic motivation learning. They also enable the robot to manage the exploration-exploitation trade-off, i.e., whether it explores further to gain more knowledge, or it exploits its internal learned model to achieve similar outcomes as humans. The probabilistic goal selection strategy also increases the sample-efficiency by selecting the most informative (novel) goals to learn from.

3. An online associative dynamic network called ”online associative radial basis function network (OARBF)” is devised to be constructed incrementally from scratch in chapter 5. This network enables Goal Babbling to learn multiple solutions of inverse models online for redundant robots. The network is updated continuously and tailored to the learned problem. To tackle the stability challenge, a parameter-sharing technique is established which synchronizes two learners online and leverages their advantages. Using the parameter sharing technique speeds up the learning process and increases sample-efficiency by drastically reducing the number of required training samples for OARBF, by drastically reducing the dimensionality of parame-

ter space, and by synchronizing two learners' updates.

Four novel learning schemes are established integrating the proposed methods:

1. Hierarchal interest-driven exploration scheme (cf. Sec. 3.1).

2. Extrinsic-intrinsic motivation learning scheme (cf. Sec. 4.1).

3. Hierarchical interest-driven associative Goal Babbling scheme (cf. Sec. 5.5).

4. Hierarchical extrinsic-intrinsic motivation-driven associative Goal Babbling scheme (cf. Appendix .4).

Each learning scheme integrates some of the proposed methods to enable different functionalities based on the required learning task. All the proposed methods and the learning schemes are data-driven, fully online, and updated on the fly. Three real-world experiments using a physical 7-DoF Baxter robot arm were conducted in order to demonstrate the applicability of the proposed methods and learning schemes. During these experiments, the robot acquired reaching skills from scratch without previous knowledge of the internal kinematic forward or inverse models, in a learning while behaving fashion, and with direct online training on a real robot without any offline training. Within a reasonable time, with reasonable accuracy, and sample-efficiency, the robot was able to:

1. self-explore a desired workspace driven by intrinsic motivation utilizing the interest measurement, and gain good knowledge about its internal inverse kinematics model during less than 4 hours of direct online training (cf. chapter 3).

2. expand its knowledge about the workspace benefiting from observing human demonstrations and achieve similar outcomes as humans within only two hours of direct training utilizing the extrinsic-intrinsic motivation learning scheme (cf. chapter 4).

3. decide autonomously on the fly on which exploration strategy to follow utilizing the novelty methods (cf. chapter 4).

4. learn two solutions to achieve the desired outcomes utilizing OARBF within 4 hours, where additional 6 hours of training were demonstrated to show the stability of OARBF utilizing the parameter-sharing technique (cf. chapter 5).

5. demonstrate smooth motion and autonomously select one of the learned configurations to achieve the task based on its previous state utilizing OARBF (cf. chapter 5).

6. decide autonomously which goal to learn, and select novel and difficult-to-attain goals to learn in order to increase sample-efficiency utilizing the interest measurement and the probabilistic goal selection strategy.

7. generalize well on new goals scattered in the learned workspace.

In all conducted real-world experiments, the robot's performance error converged very fast already after the first learning epoch, i.e., 1000 time step (samples), with an acceptable RMSE, compared to the low positioning accuracy of Baxter robot [45]. Additional training epochs were performed to enhance the performance accuracy and show the stability of the learning systems. Each epoch in the experiments took between 50 *min* to 75 *min*. The time increased based on the size of the discovered workspace as well as the complexity of the task and the learning scheme. Each learning scheme has different learning complexity based on the methods it contains. The real experiments were conducted using ROS, MATLAB, and python.

In addition, the interest measurement surpassed other state-of-the-art in terms of accuracy and robustness. The interest measurement achieved the minimal RMSE as well as the minimal RMSE std. This thesis also showed experimentally that combining knowledge-based and competence-based signals improves significantly the performance of the competence-based signals. The robot demonstrated robust performance and consistent interest despite the instantaneous updates of all the signals (interest measurement, relative error, forgetting factor, probabilistic intrinsic, and a probabilistic extrinsic signal).

Furthermore, OARBF utilizing parameter-sharing technique demonstrated high stability despite the noisy data produced from learning while behaving, the low positioning accuracy of Baxter, and the high dynamics of the network. The execution time for the iterative feedback loop to converge to one of the learned solutions was negligible. In contrast to offline ARBF, OARBF adapts its size on the fly to be tailored to the yet unknown problem, and it does not require storing any data set. OARBF also enables simultaneous exploration and solution consolidation. Thus OARBF is more compatible for lifelong learning with continuous update ability and incremental construction. Using parameter sharing technique reduces drastically the number of required training samples at least 15 times and the dimensionality of parameter space to the half.

Finally, I would like to conclude this summary by highlighting the lifelong learning components and listing how each component is achieved in this thesis:

1. Sample-Efficiency: The sample-efficiency has been increased significantly by utilizing:

1. interest measurement to select the most informative goals to learn from.
2. the probabilistic goal selection strategy to select the most novel goals to learn from.
3. OEMR to intensify the robot's experiences which consequently accelerates the convergence of the learner rapidly.
4. parameter-sharing technique to accelerate the convergence of OARBF rapidly and drastically reduces the dimensionally of parameter space .

2. Online Learning: Online learning is achieved by increasing the sample-efficiency as well as by:

1. instantaneous processing of each received sample.
2. instantaneous update for all learners, learning schemes, measures, and learning signals.

3. utilizing only the last epoch for OEMR without the need for fully storing or augmenting data.
4. instantaneous decision on which strategy to follow by utilizing the extrinsic-intrinsic learning scheme.
5. instantaneous decision on exploration-exploitation by utilizing the novelty threshold which is inferred automatically from the current robot knowledge.
6. instantaneous decision on which goal to learn by utilizing the interest measurement and the probabilistic goal selection strategy.
7. fast convergence of the feedback loop of OARBF to select one of the learned solutions with a negligible execution time.
8. learning while behaving fashion utilizing interest-driven Goal Babbling.
9. direct online training on physical robots without any offline learning.

3. Adaptability: The adaptability has been achieved by:

1. incremental learning using LLM and OARBF.
2. the learners' complexity (size) is updated on the fly to be tailored to the yet unknown problem.
3. instantaneous update for all learners, learning schemes, measures, and learning signals to adapt to any changes or newly received data.
4. fast expanding of the robot's knowledge by observing human demonstrations' outcomes, utilizing the extrinsic-intrinsic learning scheme.

4. Stability: The stability has been recognized by:

1. the high stability of OARBF without any oscillation of the performance error by utilizing the parameter-sharing technique.
2. selecting one of the learned solutions based on the previous robot state with OARBF without switching between solutions or averaging between them.
3. the consistent interest and the robust performance of the robot over all experiments despite the instantaneous update of all measure and learners.
4. the high stability of LLM despite the continuous training and exploitation during the exploration.

5. Flexibility: The flexibility has been gained by learning multiple-solutions for each required task utilizing OARBF.

6.2 Outlook

This thesis paves the way toward devising efficient and stable online learning schemes for other robot learning scenarios as well. Several avenues of potentially fruitful research could be based upon employing some of the proposed methods. Each proposed method in this thesis is independent of each other and is neither bounded to the reaching task nor tied to the proposed learning schemes. It can hence be easily integrated with other learning methods / schemes. For example, the interest measurement in chapter 3 and the probabilistic extrinsic and intrinsic signals in chapter 4 have a great potential to be integrated with the penalty signal proposed in [27] in order to detect unreachable / unlearn-able goals and to detect the robot's limits.

There are also several directions to extend the work. It will be interesting to employ self-discovery goals. For instance, direction sampling [36, 136] has been proposed as an extension of Goal Babbling. In direction sampling, the robot discovers the workspace without the need for predefined goals. This method could be adopted to allow the robot to discover goals on its own. Another possible future work derived from chapter 5 is to enable the robot to discover other solutions on its own and determine different home positions, e.g., using a specific configurations' deviation criterion. In addition, the strategy for selecting one of the learned solutions can be extended to include obstacle and self-collision avoidance. The LLM could be also extended towards an online associative LLM, and it will be interesting to compare it with OARBF in terms of accuracy and stability.

While this thesis deals particularly with learning inverse kinematics mappings, it would be also interesting to extend the work for learning sequences of behaviors and spatio-temporal patterns. Since LLM and ARBF can only learn a static mapping, learning spatio-temporal patterns could be done by using variational RNNs similar to [58] or a multi timescale MTRNN [137]. Furthermore, because Goal Babbling relies on continuous path generation between selected goals, a variational RNN could be trained in parallel to lean sequences of these goals. However, further investigation for rendering learning more efficient with variational RNN is required for direct online learning in real-world applications. In addition, the learning scheme proposed in chapter 4 could be extended to integrate other "learning from a teacher" methods, i.e., learning from demonstration and imitation learning. This could enable the robot to benefit from any information available from humans or other robots.

Appendices

.1 Local Linear Map

Local Linear Map (LLM) [44] is employed in the original Goal Babbling [44] as well as in the interest-driven Goal Babbling (cf. Sec. 3.3) in this thesis, as an incremental regression is needed for approximating the robot models online and incrementally. In principle, any regression algorithm could be used. LLM has been chosen since it has demonstrated high accuracy and stability in real robot applications (e.g, [39]) as well as for approximating complex models (e.g.,[35]).

LLM for approximating the inverse kinematics works as follows: The inverse estimate $\hat{G}(x)$ is initialized with a first local linear function $\hat{G}^{(1)}(x)$ which is centered around a prototype vector $x_p{}^{(1)} = x^{home}$, and yields the corresponding initial configuration q^{home}. M different new local linear functions $\hat{G}^{(i)}(x)$ are added incrementally during learning, centered around prototype vectors $x_p{}^{(i)}$ and active only if a new sample is received in their close vicinity which is determined by a radius r.

Let χ_i denotes a local position vector given by Eq. (1):

$$\chi_i = \left(\frac{x^* - x_p{}^{(i)}}{r} \right) \tag{1}$$

The inverse estimate $\hat{G}(x)$ is updated continuously and consists of a linear combination of the added local linear functions $\hat{G}^{(i)}(\chi_i)$, weighted by a Gaussian responsibility function $\mathrm{GR}(x)$ as given in Eq. (2).

$$\left.\begin{aligned}
\hat{G}(x^*) &= \frac{1}{N(x^*)} \sum_{i=1}^{M} \mathrm{GR}(\chi_i) \cdot \hat{G}^{(i)}(\chi_i) \\
\mathrm{GR}(\chi) &= \exp\left(- ||\chi||^2 \right) \\
N(x^*) &= \sum_{i=1}^{M} \mathrm{GR}(\chi_i) \\
\hat{G}^{(i)}(\chi_i) &= W^{(i)} \cdot \chi_i + o^{(i)},
\end{aligned}\right\} \tag{2}$$

$N(x^*)$ normalizes the Gaussian responsibility functions in the inverse estimate to scale the sum of influences of the components to unity.

The first linear function $\hat{G}^{(1)}(x)$ is initialized with $x_p{}^{(1)} = x^{home}$, $o^{(1)} = q^{home}$,

$W^{(1)} = 0$, and $\hat{G}^{(1)}(x) = q^{home}$. A new local linear function $\hat{G}^{(i+1)}(x)$ will be added when the learner receives a new training sample x_{new} at distance of at least r to all existing prototypes (i.e., $dist(x_{new}, {x_p}^{(i)}) \geqslant r$). The corresponding prototype vector is added (${x_p}^{(i+1)} = x_{new}$). The offset $o^{(i+1)}$ of $\hat{G}^{(i+1)}(x)$ is initialized with the inverse estimate before adding the new function in order to avoid abrupt changes in the inverse estimate function, i.e., the insertion of the new function will not change the local behavior of $\hat{G}(x)$ at x_{new}. The weight matrix $W^{(i+1)}$ represents the slope of the linear function after inserting the new sample:

$$\left.\begin{aligned} o^{(i+1)} &= \hat{G}(x_{new}). \\ W^{(i+1)} &= \frac{\partial \hat{G}(x)}{\partial x} = J(x) \end{aligned}\right\} \tag{3}$$

where $J(x)$ is the Jacobian matrix of the inverse estimate [32].

The parameter update ($\theta = \{W, o\}$) is done at each step using online gradient descent with learning rate η in order to minimize the weighted squared error E_t given in Eq. (5) as following:

$$\left.\begin{aligned} W_{t+1}^{(i)} &= W_t^{(i)} - \eta \cdot \frac{\partial E_t}{\partial W^{(i)}} \\ o_{t+1}^{(i)} &= o_t^{(i)} - \eta \cdot \frac{\partial E_t}{\partial o^{(i)}} \end{aligned}\right\} \tag{4}$$

$$E_t = w_t^{gb} \| q_t^+ - \hat{q}_t^+ \|^2 \tag{5}$$

Note that the execution of q_t^+ will result in x_t^+ and the corresponding configuration estimated by the learner for x_t^+ is denoted by $\hat{q}_t^+$. Hence, the goal is to minimize the error between the real and the estimated configurations in order to improve the estimation accuracy. w_t^{gb} is the sample weights (cf. Sec. 3.3), and t is the time step.

The connections between the prototypes are organized and distributed based on an Instantaneous Topological Map (ITM) described in [133] which is particularly suited to online map construction.

.2 The Calculations of the Goal Selection Probabilities in The Extrinsic Motivation Learning

Let's define an event e representing that the previously selected goal is a novel goal $e = g_{i-1} \in \mathcal{N}$. If e happened, the selection set accordingly is $\mathcal{S} = \{t, a, n\}$. If e did not happen ($\bar{e}$), the selection set accordingly is $\mathcal{S} \backslash a = \{t, n\}$. Considering Eq. (4.12) and defining j as a time step, the probability of the event e to happen (i.e., selecting a novel goal) can be calculated as follows:

$$
\begin{aligned}
&P(e_j) = P(e_j \mid e_{j-1})P(e_{j-1}) + P(e_j \mid \overline{e_{j-1}})P(\overline{e_{j-1}}) \\
&P(e_j \mid e_{j-1}) = \frac{\sum_{g_i \in \mathcal{N}} w_n(g_i)}{w_0 + w_a + \sum_{g_i \in \mathcal{N}} w_n(g_i)} \\
&P(e_j \mid \overline{e_{j-1}}) = \frac{\sum_{g_i \in \mathcal{N}} w_n(g_i)}{w_0 + \sum_{g_i \in \mathcal{N}} w_n(g_i)} \\
&P(\bar{e}) = 1 - P(e)
\end{aligned}
$$

with the initial probability $P(e_0) = 0$.

The probabilities of selecting the action a, t, and n are calculated accordingly:

$$
\begin{aligned}
&P(t) = \rho_0 = P(t \mid e)P(e) + P(t \mid \bar{e})P(\bar{e}) \\
&P(t \mid e) = P(t \mid g_{i-1} \in \mathcal{N}) = \frac{w_0}{w_0 + w_a + \sum_{g_i \in \mathcal{N}} w_n(g_i)} \\
&P(t \mid \bar{e}) = P(t \mid g_{i-1} \notin \mathcal{N}) = \frac{w_0}{w_0 + \sum_{g_i \in \mathcal{N}} w_n(g_i)}
\end{aligned}
$$

$$
\begin{aligned}
&P(a) = \rho_a = P(a \mid e)P(e) + P(a \mid \bar{e})P(\bar{e}) \\
&P(a \mid e) = P(a \mid g_{i-1} \in \mathcal{N}) = \frac{w_a}{w_0 + w_a + \sum_{g_i \in \mathcal{N}} w_n(g_i)} \\
&P(a \mid \bar{e}) = P(a \mid g_{i-1} \notin \mathcal{N}) = 0
\end{aligned}
$$

$$P(n) = P(n \mid e)P(e) + P(n \mid \bar{e})P(\bar{e})$$
$$P(n \mid e) = P(n \mid g_{i-1} \in \mathcal{N}) = \frac{\sum_{g_i \in \mathcal{N}} w_n(g_i)}{w_0 + w_a + \sum_{g_i \in \mathcal{N}} w_n(g_i)}$$
$$P(n \mid \bar{e}) = P(n \mid g_{i-1} \notin \mathcal{N}) = \frac{\sum_{g_i \in \mathcal{N}} w_n(g_i)}{w_0 + \sum_{g_i \in \mathcal{N}} w_n(g_i)}$$

The probabilities of selecting a goal from the goal sets $\mathcal{T}, \mathcal{A}, \mathcal{N}$ are:

$$P(g \in \{\mathcal{T} \backslash \mathcal{A}\}) = P(\{\mathcal{T} \backslash \mathcal{A}\} \mid t)P(t) = \frac{n_{\mathcal{T}} - n_{\mathcal{A}}}{n_{\mathcal{T}}}\rho_0$$
$$P(g \in \mathcal{A}) = P(g \in \mathcal{A} \mid a)P(a) + P(g \in \mathcal{A} \mid t)P(t) = \rho_a + \frac{n_{\mathcal{A}}}{n_{\mathcal{T}}}\rho_0$$
$$P(g \in \mathcal{N}) = P(n)$$

Where $\mathcal{T} = \mathcal{G} \setminus \mathcal{G}_{novel}$ with cardinality $n_{\mathcal{T}} = \mid \mathcal{T} \mid$, $\mathcal{A} = \mathcal{G}_a \subseteq \mathcal{G}_{train}$ with cardinality $n_{\mathcal{A}} = \mid \mathcal{A} \mid$, and $\mathcal{N} = \mathcal{G}_{novel}$, with cardinality $n_{\mathcal{N}} = \mid \mathcal{N} \mid$.

.3 Probabilistic Intrinsic Signal Evaluation

In order to evaluate the probabilistic intrinsic signal and compare it with the state-of-the-art intrinsic motivation signals, similar experiments as in Sec. 3.4 are conducted. Goal Babbling has been implemented in an illustrative 10-DoF planar manipulator with different intrinsic motivation signals: the interest signal (cf. Sec. 3.2), the probabilistic intrinsic signal (cf. Sec. 4.3.2), competence measurement [11, 18], and learning progress [19].

20 experiments have been conducted for each signal, each experiment consists of 500 epochs, and each epoch consists of 100 samples, i.e., 100 time steps. The robot starts exploring its workspace and trying to reach some predefined goals $\mathcal{G}_{train}$ to gain some knowledge about its model. The goals are selected iteratively utilizing the corresponding intrinsic motivation signal. The robot tries to reach each goal with 5 time steps (intermediate targets). The probabilistic intrinsic signal has been implemented with $\alpha = 4$ (cf. Eq. (4.14)), heuristically chosen.

As demonstrated in Table .3.1, the interest signal, as well as the probabilistic intrinsic signal, outperforms the other signals in terms of performance accuracy and robustness illustrated with the minimal std RMSE as well as the minimal validation and test RMSE. The small std indicates that all the goals are always reached within the given time frame. The high std RMSE of the other learning signals indicate that the goals on the border were not always reached as illustrated in Fig. 3.4.1.

Table .3.1: Interest-Driven Goal Babbling experimental results comparison

Goal Selection	avg. Validation RMSE $[m]$	avg. Test RMSE $[m]$	avg. RMSE std $[m]$
Interest signal	$2.7 \cdot 10^{-3}$	$1.9 \cdot 10^{-3}$	$0.8 \cdot 10^{-3}$
Probabilistic signal	$3.9 \cdot 10^{-3}$	$1.9 \cdot 10^{-3}$	$1.1 \cdot 10^{-3}$
Competence signal [18]	$5.5 \cdot 10^{-3}$	$2.3 \cdot 10^{-3}$	$1.5 \cdot 10^{-3}$
Learning signal [19]	$9 \cdot 10^{-3}$	$3.4 \cdot 10^{-3}$	$3.8 \cdot 10^{-3}$

Fig. .3.1 shows the std RMSE for the validation error over 20 experiments in the

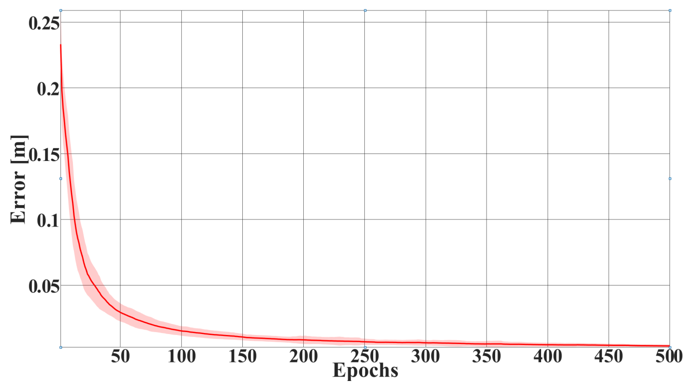


Figure .3.1: Performance error - std RMSE over 20 experiments utilizing the probabilistic intrinsic signal

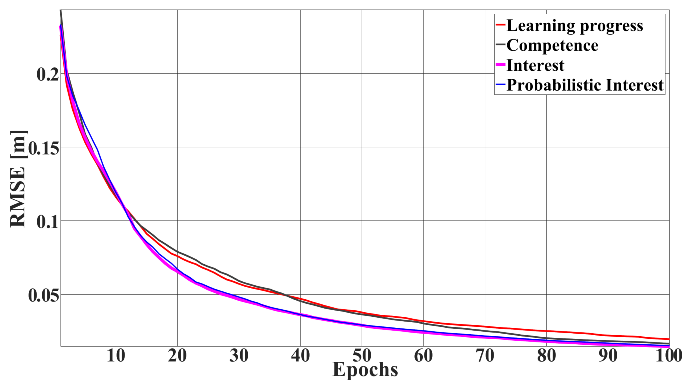
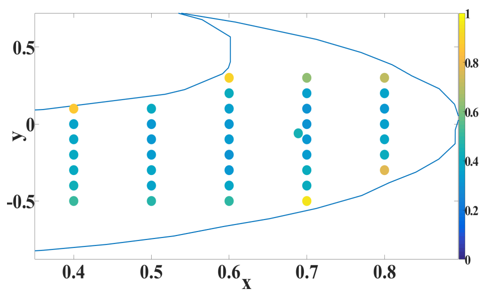


Figure .3.2: Mean performance RMSE of Goal Babbling with different intrinsic motivation signals

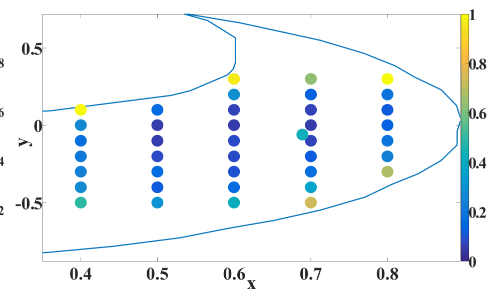
intrinsic motivation learning phase utilizing the probabilistic intrinsic signal. The error converges very fast already after 20 epochs. The performance of the robot is very robust illustrated with 1.1 mm std RMSE, which is indicated with the shaded area in Fig. .3.1.

The performance error of the interest signal, as well as the probabilistic interest signal, converges faster than the other signals as illustrated in Fig. .3.2

The median interest measurements of $\mathcal{G}_{train}$ are illustrated in Fig. .3.3 for the interest signal and the probabilistic intrinsic signal. The goals on the border are the most interesting goals for the robot, as they are difficult-to-reach indicated with the yellow dots. The starting position (home position) is illustrated with the green dot in the middle, which indicates that the robot gets interested again in the starting point due to the forgetting factor (cf. Eq. (3.2)). The main difference between these two signals is that the dark blue dotes in Fig. .3.3(b) indicate that the robot barely visited these goals and focused more on the difficult-to-reach goals. On the opposite, all goals are visited with a certain probability utilizing the probabilistic intrinsic signal, which thus can cope with the situations where there are only a few goals to learn as it can avoid getting stuck in the difficult-to-reach ones. Still, the robot managed to achieve all goals with reasonable accuracy during the exploration because the highest probabilities and the highest interests are given to the difficult-to-reach goals (cf. Fig. 3.3.1), in contrast to the signals [18, 19, 32].



(a) Probabilistic intrinsic signal



(b) Interest signal

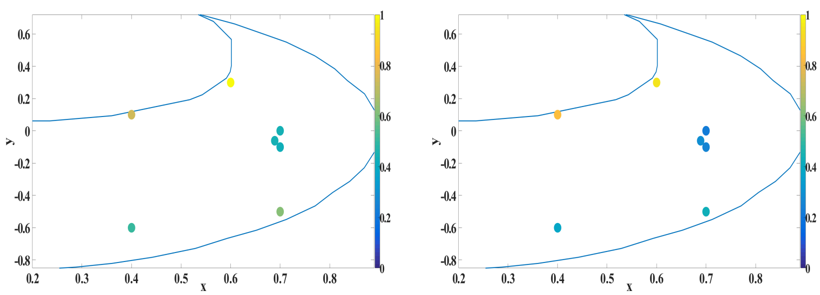
Figure .3.3: The interest measurement of the interest signal vs the probabilistic intrinsic signal

The experiments have been conducted again with only a few scattered goals to be learned in order to highlight the probabilistic intrinsic signal advantage. The goal distribution has 4 difficult-to-reach goals near the workspace border and 3 easy-to-reach goals near the home position as illustrated in Fig. .3.4. 20 experiments have been conducted for each intrinsic motivation signal with 500 epochs. As illustrated in Table .3.2, the interest signal, as well as the probabilistic intrinsic signal, outperforms the other signals in terms of accuracy and robustness. The probabilistic intrinsic signal outperforms the interest signal as it avoids get-

Table .3.2: Interest-Driven Goal Babbling experimental results comparison 2

Goal Selection	avg. Validation RMSE $[m]$	avg. Test RMSE $[m]$	avg. RMSE std $[m]$
Interest signal	$9.3 \cdot 10^{-3}$	$3.5 \cdot 10^{-3}$	$9.2 \cdot 10^{-3}$
Probabilistic signal	$5.4 \cdot 10^{-3}$	$2.3 \cdot 10^{-3}$	$3.5 \cdot 10^{-3}$
Competence signal [18]	$17 \cdot 10^{-3}$	$4.2 \cdot 10^{-3}$	$20 \cdot 10^{-3}$
Learning signal [19]	$17 \cdot 10^{-3}$	$6.9 \cdot 10^{-3}$	$16 \cdot 10^{-3}$

ting stuck in the difficult-to-reach goals. Visiting all scattered goals with the focus on the difficult-to-reach goals assures discovering the full detected workspace between these goals as well as learning the border of the workspace.



(a) Probabilistic intrinsic signal (b) Interest signal

Figure .3.4: The interest measurement of the interest signal vs the probabilistic intrinsic signal - second experiment

The median interest measurements of $\mathcal{G}_{train}$ are illustrated in Fig. .3.4 for the interest signal and the probabilistic intrinsic signal. The interest signal focused more on the difficult-to-reach goals, which are the most visited goals during the training as indicated in yellow dots. The easy-to-reach goals (the blue and green dots in the middle) are visited more utilizing the probabilistic intrinsic signal, which helps to learn better the full detected workspace. Still, the difficult-to-reach goals are the most interesting to the robot utilizing both signals.

Note that in both experiments, the test RMSE is less than the validation RMSE, due to the different distributions of the goal sets. The test goals are scattered inside the workspace, and the validation set contains a difficult-to-reach goals.

Comparing the interest focus of the interest and the probabilistic intrinsic signals in Fig. 3.3.1 and Fig. .3.4, the general behavior of both signals is always the same. Both signals give the highest interest for the difficult-to-reach goals. The main difference is that the probabilistic intrinsic signal avoids getting stuck in local minimum or local maximum as all goals are visited with a certain probability.

.4 Hierarchical Extrinsic-Intrinsic Motivation-Driven Associative Goal Babbling

It is straight forward to combine OARBF (cf. chapter 5) with the extrinsic-intrinsic learning scheme (cf. chapter 4). Fig. .4.1 illustrates the architecture of the hierarchical extrinsic-intrinsic motivation-driven associative Goal Babbling. It comprises two high levels of exploration strategy and goal selection strategy, and two low levels of exploration mechanism for incremental approximating the underlying model and associative memory for consolidating different solutions:

(A) Exploration strategy: This strategy determines whether the exploration is guided by the intrinsic or the extrinsic motivation.

(B) Goal selection strategy: When novel goals are detected from observing human demonstrations, the goals are selected utilizing the extrinsic signal (cf. Sec. 4.3.1). Otherwise, the goals are selected utilizing the intrinsic signal (cf. Sec. 4.3.2).

(C) Goal-directed exploration mechanism: The exploration relies on the interest-driven Goal Babbling (cf. Sec. 3.3) to obtain the internal model in a learning while behaving fashion.

(D) Associative dynamic memory: OARBF is implemented for online learning and consolidating multiple solutions of inverse models with Goal Babbling for redundant robots.

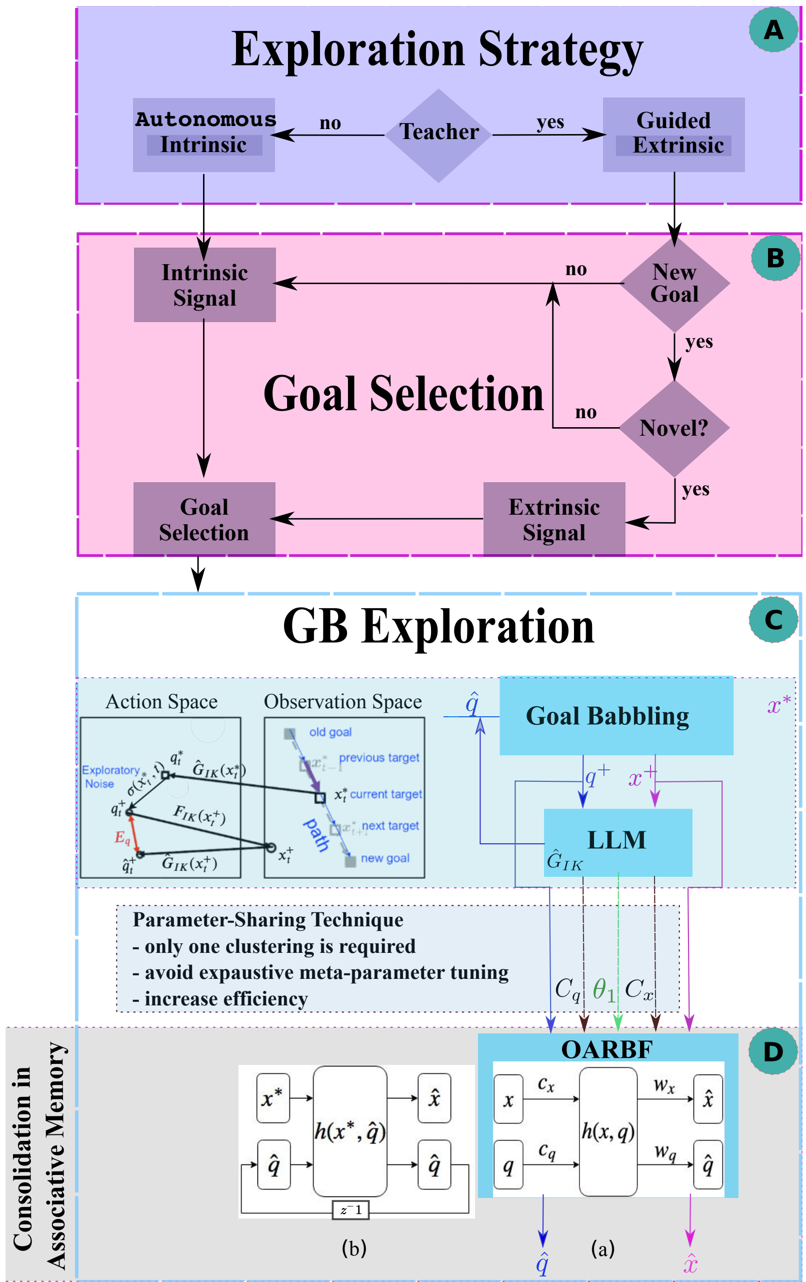


Figure .4.1: Hierarchical extrinsic-intrinsic motivation-driven associative Goal Babbling scheme. It comprises two high levels of exploration strategy and goal selection strategy, and two low levels of exploration and solution consolidation. (A) Exploration strategy, (B) Goal selection strategy, (C) Goal-directed exploration mechanism, (D): Associative dynamic memory (a) OARBF training, (b) establishing a feedback loop for OARBF exploitation. $\theta_1 = \{\eta, r\}$

Related References by the Author

[132] **R. Rayyes and J. Steil, "Online associative multi-stage goal babbling toward versatile learning of sensorimotor skills," in 2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2019, pp. 327–334.**

In this paper, I devised the online associate radial basis function network and established the parameter-sharing technique. I evaluated the methods in simulation and compared the network to the offline version.

[31] **R. Rayyes, H. Donat, and J. Steil, "Hierarchical interest-driven goal babbling for efficient bootstrapping of sensorimotor skills," in 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020, pp. 1336–1342.**

In this paper, I devised the interest measurement, relative error, and forgetting factor. I developed the interest-driven Goal Babbling and designed the hierarchical interest-driven associative goal babbling scheme. I proposed the episodic mental replay. I also evaluated the proposed methods with the online associative radial basis function network in a real robot experiment.

[125] **R. Rayyes, H. Donat, and J. Steil, "Efficient online interest-driven exploration for developmental robots", IEEE Trans. Cognitive and Developmental Systems, 2020.**

This paper extended the previous work presented in [31]. I evaluated the interest measurement in comparison to state-of-the-art methods. I also designed the hierarchical interest-driven exploration scheme

[130] **R. Rayyes, H. Donat, J. Steil, and M. Spranger, "Interest-driven exploration with observational learning for developmental robots," IEEE Trans. Cognitive and Developmental Systems, in press.**

In this paper, I designed the extrinsic-intrinsic learning scheme which combines intrinsic motivation with learning from observation. I developed three new methods for achieving observational learning and evaluated the proposed framework in simulation and on a real robot.

Publication list during my Ph.D.:

- R. Rayyes, H. Donat, J. Steil, and M. Spranger, “Interest-driven exploration with observational learning for developmental robots,” IEEE Trans. Cognitive and Developmental Systems, in press.

- R. Rayyes, H. Donat, and J. Steil, ”Efficient online interest-driven exploration for developmental robots”, IEEE Trans. Cognitive and Developmental Systems, 2020.

- R. Rayyes, H. Donat, and J. Steil, “Hierarchical interest-driven goal babbling for efficient bootstrapping of sensorimotor skills,” in 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020, pp. 1336–1342, video: https://www.rob.cs.tu-bs.de/node/809

- R. Rayyes and J. Steil, “Online associative multi-stage goal babbling toward versatile learning of sensorimotor skills,” in 2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), 2019, pp. 327–334.

- R. Rayyes, D. Kubus, and J. Steil, “Learning inverse statics models efficiently with Symmetry-based exploration”, Frontiers in Neurorobotics Journal, Vol. 12, pp. 68, 2018.

- D. Kubus, R. Rayyes, and J. J. Steil, J., ”Learning forward and inverse kinematics maps efficiently,” in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 2018, pp. 5133–5140.

- R. Rayyes, D. Kubus, and J. Steil, “Multi-stage goal babbling for learning inverse models simultaneously,” in IROS Workshop, BODIS: The Utility of Body, Interaction and Self Learning in Robotics Workshop, 2018.

- R. Rayyes, D. Kubus, C. Hartmann, and J. Steil, “Learning inverse statics models efficiently”, arxiv, 2017.

- R. Rayyes and J. J. Steil, “Goal babbling with direction sampling for simultaneous exploration and learning of inverse kinematics of a humanoid robot,” in Proc. of the WS on NC2, vol. 4, 2016, pp. 56–63.

The work has been presented at the following workshops

- R. Rayyes and J. Steil, "Hierarchal interest-driven associative goal babbling", The Annual Conference on Neural Information Processing Systems (NeurIPS), WiML workshop, Vancouver, Canada, 2019.

- R. Rayyes and J. Steil, "Interest-based exploration with associative goal babbling toward versatile cognitive robots", Robotics Science and Systems (RSS): WiR workshop, Freiburg, Germany, 2019.

- R. Rayyes and J. Steil, "Interest-based exploration with associative goal babbling", The Fourth International Workshop on Intrinsically Motivated Open-ended Learning (IMOL), Frankfurt, Germany, 2019.

- R. Rayyes and J. Steil, "Multi-stage goal babbling for learning inverse models simultaneously", NeurIPS: WiML workshop, Montreal, Canada, 2018

- D. Kubus, R. Rayyes, and J. J. Steil, J., "Learning inverse statics with online Goal Babbling", DGR-Tage workshop, Bremen, Germany, 2017.

- R. Rayyes, D. Kubus, and J. Steil, "Learning gravity compensation with online Goal Babbling", The Third International Workshop on Intrinsically Motivated Open-ended Learning (IMOL), Rome, Italy, 2017.

Scholarships and grants during my Ph.D.

- ANITA B.ORG GHC Student Scholarship for Virtual Grace Hopper Celebration (2020).

- Google award for attending womENcourage conference, Rome, Itlay, (2019).

- Nominated by DAAD for the 7th Heidelberg Laureate Forum, Heidelberg, Germany (2019).

- IAESTE award for exchange students (2019).

- DAAD travel grant (2019).

- Women in Machine Learning (WiML) travel grant (2019).

- Women in Machine Learning (WiML) travel grant (2018).
- DAAD scholarship "Research Grants – Doctoral Programme in Germany" (2015).

Ph.D. requirements and other activities

- I had to take two courses and exams as a prerequisite for a Ph.D. admission at TU Braunschweig during the winter semester 2017/2018:

- Theoretische Informatik 1 (5 credit points), with Prof. Dr. Roland Meyer.
- Grundlagen Maschinelles Lernen (5 credit points), with Prof. Dr. Jochen Steil.

- I had also to participate in the DAAD program as one of the requirements for the scholarship:

- E-Learning and workshops about advanced studies in Sustainable Economic and Personal Competence - Konstanz University (11/2016 – 8/2017).

- I joined Google Get Ahead Program in summer 2019. The program included technical challenges and trainings for selected Computer Science students from Europe, Middle East and Africa (EMEA).

- I did a research internship at Sony Computer Science Lab (CSL) Tokyo - Japan for five months (10/2019 – 03/2020).

References

[1] M. Asada, K. F. MacDorman, H. Ishiguro, and Y. Kuniyoshi, "Cognitive developmental robotics as a new paradigm for the design of humanoid robots," *Elsevier Science,Robotics and Autonomous Systems*, vol. 37, pp. 185–193, 2001.

[2] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida, "Cognitive developmental robotics: A survey," *IEEE transactions on autonomous mental development*, vol. 1, no. 1, pp. 12–34, 2009.

[3] A. Cangelosi, M. Schlesinger, and L. B. Smith, "Developmental robotics: From babies to robots," *MIT Press*, 2015.

[4] M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini, "Developmental robotics: a survey," *Connection Science*, vol. 15, no. 4, pp. 151–190, 2003.

[5] J. Schmidhuber, "Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts," *Connection Science*, vol. 18, no. 2, pp. 173–187, 2006.

[6] A. Cangelosi, G. Metta, G. Sagerer, S. Nolfi, C. Nehaniv, K. Fischer, J. Tani, T. Belpaeme, G. Sandini, F. Nori, L. Fadiga, B. Wrede, K. Rohlfing, E. Tuci, K. Dautenhahn, J. Saunders, and A. Zeschel, "Integration of action and language knowledge: A roadmap for developmental robotics," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 167–195, 2010.

[7] H. Kim, H. Jasso, G. Deak, and J. Triesch, "A robotic model of the development of gaze following," in *2008 7th IEEE International Conference on Development and Learning*, 2008, pp. 238–243.

[8] N. S. Mai, “A curious robot learner for interactive goal-babbling: : Strategically choosing what, how, when and from whom to learn,” *Universite de Bordeaux*, 2013.

[9] E. Ugur, Y. Nagai, E. Sahin, and E. Oztop, “Staged development of robot skills: Behavior formation, affordance learning and imitation with motionese,” *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 2, pp. 119–139, June 2015.

[10] S. Forestier, “Intrinsically motivated goal exploration in child development and artificial intelligence : Learning and development of speech and tool use,” *Université Bordeaux*, 2019.

[11] S. M. Nguyen and P. Oudeyer, “Socially guided intrinsic motivation for robot learning of motor skills,” *Autonomous Robots*, vol. abs/1804.07269, 2014.

[12] D. Nguyen-Tuong and J. Peters, “Model learning for robot control: a survey,” *Cognitive Processing*, vol. 12, no. 4, pp. 319–340, Apr. 2011.

[13] S. Forestier and P. Oudeyer, “Modular active curiosity-driven discovery of tool use,” in *IEEE/RSJ, IROS*, 2016, pp. 3965–3972.

[14] M. Rolf, J. J. Steil, and M. Gienger, “Learning flexible full body kinematics for humanoid tool use,” in *2010 International Conference on Emerging Security Technologies*. IEEE, 2010, pp. 171–176.

[15] X. Huang and J. Weng, “Motivational system for human-robot interaction,” in *Computer Vision in Human-Computer Interaction*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 17–27.

[16] F. de La Bourdonnaye, C. Teulière, J. Triesch, and T. Chateau, “Stage-wise learning of reaching using little prior knowledge,” *Frontiers in Robotics and AI*, vol. 5, p. 110, 2018.

[17] G. Baldassarre, “Intrinsic motivations and open-ended learning,” *arXiv*, 2019.

[18] A. Baranes and P. Oudeyer, “Active learning of inverse models with intrinsically motivated goal exploration in robots,” *Robot. Auton. Syst.*, vol. 61, no. 1, pp. 49–73, 2013.

[19] V. G. Santucci, G. Baldassarre, and M. Mirolli, "GRAIL: A goal-discovering robotic architecture for intrinsically-motivated learning," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8, no. 3, pp. 214–231, Sep. 2016.

[20] J. Schmidhuber, "Formal theory of creativity, fun, and intrinsic motivation (1990 - 2010)," *IEEE Trans. on Auton. Ment. Dev.*, vol. 2, no. 3, pp. 230–247, Sep. 2010.

[21] D. Borsa, N. Heess, B. Piot, S. Liu, L. Hasenclever, R. Munos, and O. Pietquin, "Observational learning by reinforcement learning," in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, ser. AAMAS '19. International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1117–1124.

[22] A. Bandura and R. Walters, *Social learning and personality development*. Holt, Rinehart and Winston, 1963.

[23] A. Bandura, *Social learning theory*. Prentice-Hall, 1977.

[24] M. Frank, J. Leitner, M. Stollenga, A. Förster, and J. Schmidhuber, "Curiosity driven reinforcement learning for motion planning on humanoids," *Frontiers in Neurorobotics*, vol. 7, p. 25, 2014.

[25] D. Tanneberg, J. Peters, and E. Rueckert, "Intrinsic motivation and mental replay enable efficient online adaptation in stochastic recurrent networks," *CoRR*, vol. abs/1802.08013, 2018.

[26] P. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 2, pp. 265–286, April 2007.

[27] S. H. Huang, M. Zambelli, J. Kay, M. F. Martins, Y. Tassa, P. M. Pilarski, and R. Hadsell, "Learning gentle object manipulation with curiosity-driven deep reinforcement learning," *CoRR*, vol. abs/1903.08542, 2019.

[28] S. Forestier, R. Portelas, Y. Mollard, and P.-Y. Oudeyer, "Intrinsically Motivated Goal Exploration Processes with Automatic Curriculum Learning," *arXiv e-prints*, p. arXiv:1708.02190, Aug. 2017.

[29] N. Duminy, S. M. Nguyen, and D. Duhaut, "Strategic and interactive learning of a hierarchical set of tasks by the poppy humanoid robot," in *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2016, pp. 204–209.

[30] D. Kubus, R. Rayyes, and J. J. Steil, "Learning forward and inverse kinematics maps efficiently," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain, 2018, pp. 5133–5140.

[31] R. Rayyes, H. Donat, and J. Steil, "Hierarchical interest-driven goal babbling for efficient bootstrapping of sensorimotor skills," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 1336–1342.

[32] M. Rolf, J. J. Steil, and M. Gienger, "Online goal babbling for rapid bootstrapping of inverse models in high dimensions," in *IEEE Int. Conf. Development and Learning and on Epigenetic Robotics*, 2011, pp. 1–8.

[33] C. von Hofsten, "An action perspective on motor development," *Trends in CogSci*, vol. 8, p. 266–272, 2004.

[34] A. K. Philippsen, R. F. Reinhart, and B. Wrede, "Goal babbling of acoustic-articulatory models with adaptive exploration noise," in *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 2016, pp. 72–78.

[35] R. Rayyes, D. Kubus, and J. Steil, "Learning inverse statics models efficiently with symmetry-based exploration," *Frontiers in Neurorobotics*, vol. 12, p. 68, 2018.

[36] R. Rayyes and J. J. Steil, "Goal babbling with direction sampling for simultaneous exploration and learning of inverse kinematics of a humanoid robot," in *Proc. of the WS on NC2*, vol. 4, 2016, pp. 56–63.

[37] R. Rayyes, D. Kubus, and J. Steil, "Multi-stage goal babbling for learning inverse models simultaneously," in *IROS Workshop, BODIS: The Utility of Body, Interaction and Self Learning in Robotics Workshop*, 2018.

[38] F. Gama, M. Shcherban, M. Rolf, and M. Hoffmann, "Active exploration for body model learning through self-touch on a humanoid robot with artificial skin," in *2020 Joint IEEE 10th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2020, pp. 1–8.

[39] M. Rolf and J. Steil, “Efficient exploratory learning of inverse kinematics on a bionic elephant trunk,” in *IEEE Trans. Neural Networks and Learning Systems*, vol. 25, no. 6, 2014, pp. 1147–1160.

[40] R. Reinhart and M. Rolf, “Learning versatile sensorimotor coordination with goal babbling and neural associative dynamics,” in *IEEE ICDL*, Aug 2013, pp. 1–7.

[41] R. F. Reinhart and J. J. Steil, “Learning whole upper body control with dynamic redundancy resolution in coupled associative radial basis function networks,” in *IROS*. IEEE, 2012, pp. 1487–1492.

[42] R. Reinhart, *Reservoir computing with output feedback*. Bielefeld University, 2011.

[43] G. Parisi, R. Kemker, J. Part, C. Kanan, and S. Wermter, “Continual lifelong learning with neural networks: A review,” *Neural Networks*, vol. 113, pp. 54–71, May 2019.

[44] H. Ritter, “Learning with the Self-Organizing Map,” in *ICANN-91*, T. Kohonen, Ed., vol. 1. North Holland, 1991, pp. 379–384.

[45] “Hardware specifications,” http://sdk.rethinkrobotics.com/wiki/Hardware_Specifications, accessed: 2019-09-12.

[46] N. H. Siddique, P. Dhakan, I. Rañó, and K. E. Merrick, “A review of the relationship between novelty, intrinsic motivation and reinforcement learning,” *Paladyn, Journal of Behavioral Robotics*, vol. 8, pp. 58–69, 2017.

[47] A. F. Baranes, P.-Y. Oudeyer, and J. Gottlieb, “The effects of task difficulty, novelty and the size of the search space on intrinsically motivated exploration,” *Frontiers in Neuroscience*, vol. 8, p. 317, 2014.

[48] P.-Y. Oudeyer and F. Kaplan, “What is intrinsic motivation? a typology of computational approaches,” *Frontiers in Neurorobotics*, vol. 1, p. 6, 2007.

[49] V. Santucci, G. Baldassarre, and M. Mirolli, “Which is the best intrinsic motivation signal for learning multiple skills?” *Frontiers in Neurorobotics*, vol. 7, p. 22, 2013.

[50] A. Barto, M. Mirolli, and G. Baldassarre, “Novelty or surprise?” *Frontiers in Psychology*, vol. 4, p. 907, 2013.

[51] F. Benureau and P.-Y. Oudeyer, "Behavioral diversity generation in autonomous exploration through reuse of past experience," *Frontiers in Robotics and AI*, vol. 3, p. 8, 2016.

[52] C. Zhang, Y. Zhao, J. Triesch, and B. E. Shi, "Intrinsically motivated learning of visual motion perception and smooth pursuit," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 1902–1908.

[53] N. Chentanez, A. G. Barto, and S. P. Singh, "Intrinsically motivated reinforcement learning," in *Advances in Neural Information Processing Systems 17*. MIT Press, 2005, pp. 1281–1288.

[54] J. Metzen and F. Kirchner, "Incremental learning of skill collections based on intrinsic motivation," *Frontiers in Neurorobotics*, vol. 7, p. 11, 2013.

[55] J. Storck, S. Hochreiter, and J. Schmidhuber, "Reinforcement driven information acquisition in non-deterministic environments," *ICANN*, pp. 159 – 164, 1994.

[56] P. Schwartenbeck, T. Fitzgerald, R. J. Dolan, and K. Friston, "Exploration, novelty, surprise, and free energy minimization," *Frontiers in psychology*, vol. 4, no. 710, pp. 1–15, 2013.

[57] R. Kaplan and K. J. Friston, "Planning and navigation as active inference," *Biological Cybernetics*, vol. 112, pp. 323 – 343, 2018.

[58] A. Ahmadi and J. Tani, "A novel predictive-coding-inspired variational RNN model for online prediction and recognition," *Neural Computation*, vol. 31, no. 11, 2019.

[59] D. Caligiore1, D. C. Magda Mustile1, P. Redgrave, J. Triesch, M. D. Marsico, and G. Baldassarre, "Intrinsic motivations drive learning of eye movements: An experiment with human adults," *PLOS ONE*, vol. 10, no. 3, pp. 1–15, 03 2015.

[60] A. Barto, S. Singh, and N. Chentanez, "Intrinsically motivated learning of hierarchical collections of skills," in *ICDL*, 2004.

[61] J. Schmidhuber, "Curious model-building control systems," *IEEE International Joint Conference on Neural Networks*, p. 1458–1463, 1991a.

[62] M. Markou and S. Singh, "Novelty detection: A review - part 2: Neural network based approaches," *Signal Processing*, vol. 83, pp. 2499–2521, 2003.

[63] M. Markou and S. Singh, "Novelty detection: A review - part 1: Statistical approaches," *Signal Processing*, vol. 83, pp. 2481–2497, 2003.

[64] S. Hart and R. Grupen, "Learning generalizable control programs," *IEEE Transactions on Autonomous Mental Development*, vol. 3, no. 3, pp. 216–231, 2011.

[65] D. Foster and M. Wilson, "Reverse replay of behavioural sequences in hippocampal place cells during the awake state," *Nature*, vol. 440, no. 7084, pp. 680–683, 3 2006.

[66] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[67] L.-J. Lin, *Reinforcement learning for robots using neural networks*. Technical report, DTIC Document, 1993.

[68] A. Gerken and M. Spranger, "Continuous value iteration (CVI) reinforcement learning and imaginary experience replay (IER) for learning multi-goal, continuous action and state space controllers," *IEEE ICRA*, 2019.

[69] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, P. Abbeel, and W. Zaremba, "Hindsight experience replay," *Advances in Neural Information Processing Systems 30*, pp. 5048–5058, 2017.

[70] M. Riedmiller, R. Hafner, T. Lampe, M. Neunert, J. Degrave, T. van de Wiele, V. Mnih, N. Heess, and J. T. Springenberg, "Learning by playing solving sparse reward tasks from scratch," ser. Proceedings of Machine Learning Research, vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 4344–4353.

[71] D. Zhao, Haitao Wang, Kun Shao, and Y. Zhu, "Deep reinforcement learning with experience replay based on sarsa," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2016, pp. 1–6.

[72] M. Fang, T. Zhou, Y. Du, L. Han, and Z. Zhang, "Curriculum-guided hindsight experience replay," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 12 623–12 634.

[73] R. Zhao and V. Tresp, "Energy-based hindsight experience prioritization," *CoRL*, pp. 113–122, 2018.

[74] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," *arXiv*, 2015.

[75] J. Luo and H. Li, "Dynamic experience replay," *arXiv*, 2020.

[76] Y. Lin, J. Huang, M. Zimmer, Y. Guan, J. Rojas, and P. Weng, "Invariant transform experience replay: Data augmentation for deep reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6615–6622, 2020.

[77] G. I. Parisi, J. Tani, C. Weber, and S. Wermter, "Lifelong learning of spatiotemporal representations with dual-memory recurrent self-organization," *Frontiers in Neurorobotics*, vol. 12, p. 78, 2018.

[78] G. van de Ven, H. Siegelmann, and A. Tolias, "Brain-inspired replay for continual learning with artificial neural networks," *Nature Communications*, 2020.

[79] L. S. Vygotskii and M. Cole, *Mind in society : the development of higher psychological processes / L. S. Vygotsky*. Harvard University Press Cambridge, 1978.

[80] J. Ho and S. Ermon, "Generative adversarial imitation learning," in *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 2016, pp. 4565–4573.

[81] F. Torabi, G. Warnell, and P. Stone, "Behavioral cloning from observation," in *IJCAI*, ser. IJCAI'18. AAAI Press, 2018, pp. 4950–4957.

[82] A.-L. Vollmer, M. Mühlig, J. J. Steil, K. Pitsch, J. Fritsch, K. Rohlfing, and B. Wrede, "Robots show us how to teach them: Feedback from robots shapes tutoring behavior during action learning," *PLoS ONE*, p. 1159–1169, 2014.

[83] C. Tennie, J. Call, and M. Tomasello, "Push or pull: Imitation vs. emulation in great apes and human children," *Ethology*, vol. 112, p. 1159–1169, 2006.

[84] J. Triesch, "Imitation learning based on an intrinsic motivation mechanism for efficient coding," in *Frontiers in psychology*, vol. 4, no. 800, 2013.

[85] A. L. Thomaz and C. Breazeal, "Robot learning via socially guided exploration," in *2007 IEEE 6th International Conference on Development and Learning*, July 2007, pp. 82–87.

[86] N. Duminy, S. M. Nguyen, and D. Duhaut, "Learning a set of interrelated tasks by using a succession of motor policies for a socially guided intrinsically motivated learner," *Frontiers in Neurorobotics*, vol. 12, p. 87, 2019.

[87] O. L. Georgeon, J. B. Marshall, and S. Gay, "Interactional motivation in artificial systems: Between extrinsic and intrinsic motivation," in *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, 2012, pp. 1–2.

[88] R. M. Ryan and E. L. Deci, "Intrinsic and extrinsic motivations: Classic definitions and new directions," *Contemporary Educational Psychology*, vol. 25, no. 1, pp. 54 – 67, 2000.

[89] L. Lonini, S. Forestier, C. Teulière, Y. Zhao, B. Shi, and J. Triesch, "Robust active binocular vision through intrinsically motivated learning," *Frontiers in Neurorobotics*, vol. 7, p. 20, 2013.

[90] M. R. Lepper and D. I. Cordova, "A desire to be taught: Instructional consequences of intrinsic motivation," *Motivation and Emotion*, vol. 16, no. 3, pp. 187–208, Sep 1992.

[91] D. Wolpert and M. Kawato, "Multiple paired forward and inverse models for motor control," *Neural Networks 1*, vol. 11, p. 1317–1329, 1998.

[92] Y. Demiris and A. Meltzoff, "The robot in the crib: A developmental analysis of imitation skills in infants and robots," in *Infant and child development*, vol. 17, no. 1, 2008, pp. 43–53.

[93] A. Dearden and Y. Demiris, "Learning forward models for robots," in *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, ser. IJCAI'05. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005, pp. 1440–1445.

[94] pathway.org, "The 4 to 6 month baby communication milestones to look for," https://www.youtube.com/watch?v=doFGHFrMRXI, 2016 Accessed: 2019-09-14.

[95] M. Rolf, J. Steil, and M. Gienger, "Goal babbling permits direct learning of inverse kinematics," *IEEE Trans. Autonomous Mental Development*, vol. 2, pp. 216–229, 2010.

[96] A. Baranes and P. Oudeyer, "Intrinsically motivated goal exploration for active motor learning in robots: A case study," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 1766–1773.

[97] P. Loviken, N. Hemion, A. Laflaquière, M. Spranger, and A. Cangelosi, "Online learning of body orientation control on a humanoid robot using finite element goal babbling," in *IROS*. IEEE, 2018.

[98] R. F. Reinhart, "Autonomous exploration of motor skills by skill babbling," *Auton. Robots*, vol. 41, no. 7, pp. 1521–1537, 2017.

[99] M. Schmerling, G. Schillaci, and V. Hafner, in *Goal-directed learning of hand-eye coordination in a humanoid robot*, 2015, p. 68–175.

[100] M. Ito and J. Tani, "On-line imitative interaction with a humanoid robot using a mirror neuron model," *IEEE International Conference on Robotics and Automation(ICRA)*, vol. 2, pp. 1071–1076, 2004.

[101] P. Pastor, H. Hoffmann, T. Asfour, and S. Schaal, "Learning and generalization of motor skills by learning from demonstration," *2009 IEEE International Conference on Robotics and Automation*, pp. 763–768, 2009.

[102] S. Schaal, A. Ijspeert, and A. Billard, "Computational approaches to motor learning by imitation," *Philosophical Transaction of the Royal Society of London: Series B, Biological Sciences*, vol. 358, no. 1431, pp. 537–547, 2003, clmc.

[103] A. J. Ijspeert, J. Nakanishi, and S. Schaal, "Movement imitation with nonlinear dynamical systems in humanoid robots," in *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292)*, vol. 2, 2002, pp. 1398–1403 vol.2.

[104] R. F. Reinhart and J. J. Steil, "Reaching movement generation with a recurrent neural network based on learning inverse kinematics for the humanoid robot icub," *2009 9th IEEE-RAS International Conference on Humanoid Robots*, pp. 323–330, 2009.

[105] R. F. Reinhart and J. Steil, "Attractor-based computation with reservoirs for online learning of inverse kinematics," *ESANN*, vol. 2, pp. 257–262, 2009.

[106] M. Jordan and D. Rumelhart, "Forward models: supervised learning with a distal teacher," *CognSci*, vol. 16, pp. 307–354, 1992.

[107] R. F. Reinhart and J. J. Steil, "Neural learning and dynamical selection of redundant solutions for inverse kinematic control," in *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2011, pp. 564–569.

[108] J. Barhen, S. Gulati, and M. Zak, "Neutral learning of constrained nonlinear transformations," *Computer*, vol. 22, no. 6, pp. 67–76, 1989.

[109] M. Lukosevicius and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," *Comput. Sci. Rev.*, vol. 3, pp. 127–149, 2009.

[110] G. bin Huang, Q. yu Zhu, and C. kheong Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, pp. 489–501, 2006.

[111] J. Tani, M. Ito, and Y. Sugita, "Self-organization of distributedly represented multiple behavior schemata in a mirror system: reviews of robot experiments using RNNPB," *Neural networks : the official journal of the International Neural Network Society*, vol. 17, no. 8-9, p. 1273—1289, 2004.

[112] I. Igari and J. Tani, "Incremental learning of sequence patterns with a modular network model," *Neurocomputing*, vol. 72, pp. 1910–1919, 2009.

[113] R. Reinhart and J. Steil, "State prediction: A constructive method to program recurrent neural networks," in *Artificial Neural Networks and Machine Learning – ICANN*, vol. 6791. Springer, Berlin, Heidelberg, 2011.

[114] Y. Tomikawa and K. Nakayama, "Approximating many valued mappings using a recurrent neural network," *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227)*, vol. 2, pp. 1494–1497 vol.2, 1998.

[115] M. Lukosevicius, "On self-organizing reservoirs and their hierarchies," *Jacobs University Technical Reports*, no. 25, 2010.

[116] T. Voegtlin and P. Dominey, "Recursive self-organizing maps," *Advances in Self-Organising Maps*, 2001.

[117] D. Broomhead and D. Lowe, "Multivariable functional interpolation and adaptive networks," *Complex Systems*, vol. 2, pp. 321–355, 1988.

[118] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, pp. 79–87, 1991.

[119] D. Koert, G. Maeda, G. Neumann, and J. Pcters, "Learning coupled forward-inverse models with combined prediction errors," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 2433–2439.

[120] B. Damas and J. Santos-Victor, "An online algorithm for simultaneously learning forward and inverse kinematics," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 321–355.

[121] C. M. Bishop, "Mixture density networks," Tech. Rep., 1994.

[122] D'Souza, S. Vijayakumar, and S. Schaal, "Learning inverse kinematics," *Int. Conf. Intelligent Robots and Systems (IROS)*, vol. 1, pp. 298 – 303, 2001.

[123] B. Bócsi, D. Nguyen-Tuong, L. Csató, B. Schölkopf, and J. Peters, "Learning inverse kinematics with structured prediction," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 698–703.

[124] T. Matsumoto and J. Tani, "Goal-directed planning for habituated agents by active inference using a variational recurrent neural network," *Entropy*, vol. 22, no. 5, 2020.

[125] R. Rayyes, H. Donat, and J. Steil, "Efficient online interest-driven exploration for developmental robots," *IEEE Trans. Cognitive and Developmental Systems*, 2020.

[126] P. O. Stalph and M. V. Butz, "Learning local linear jacobians for flexible and adaptive robot arm control," *Genetic Programming and Evolvable Machines*, vol. 13, no. 2, pp. 137–157, 2012.

[127] M. O'Mahony, *Sensory Evaluation of Food: Statistical Methods and Procedures*. CRC Press, 2017.

[128] R. Rayyes, H. Donat, and J. Steil, "Interest-driven (associative) goal babbling with Baxter," https://https://www.rob.cs.tu-bs.de/node/809/.

[129] "Moveit," https://moveit.ros.org/, accessed: 2019-09-14.

[130] R. Rayyes, H. Donat, J. Steil, and M. Spranger, "Interest-driven exploration with observational learning for developmental robots," *IEEE Trans. Cognitive and Developmental Systems*, press.

[131] "Statistics: Power from data, statistics canada, canada, 2017," https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch12/5214876-eng.htm, accessed: 2020-07-10.

[132] R. Rayyes and J. Steil, "Online associative multi-stage goal babbling toward versatile learning of sensorimotor skills," in *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2019, pp. 327–334.

[133] J. Jockusch and H. Ritter, "An instantaneous topological mapping model for correlated stimuli," in *Neural Networks, 1999. IJCNN '99. International Joint Conference on*, vol. 1, 1999, pp. 529–534.

[134] M. Rolf, "Goal babbling for an efficient bootstrapping of inverse models in high dimensions." Bielefeld University, 2012.

[135] R. Lewis and V. Torczon, "Pattern search algorithms for bound constrained minimization," *SIAM Journal on Optimization*, vol. 9, no. 4, pp. 1082–1099, 1999.

[136] M. Rolf, "Goal babbling with unknown ranges: A direction-sampling approach," in *IEEE Int. Conf. on Development and Learning and on Epigenetic Robotics (ICDL)*, 2013, pp. 1–7.

[137] Y. Yamashita and J. Tani, "Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment," in *PLoS computational biology*, vol. 4. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, p. 11.